8 29°
6.8.70

Vol. VIII Part I May 1955

THE BRITISH JOURNAL OF STATISTICAL PSYCHOLOGY

EDITED BY
CYRIL BURT

WITH THE ASSISTANCE OF
CHARLOTTE BANKS AND ALAN STUART
AND THE FOLLOWING EDITORIAL BOARD

A. C. AITKEN
M. S. BARTLETT
W. G. EMMETT
M. G. KENDALL
D. N. LAWLEY
E. S. PEARSON
E. A. PEEL
L. S. PENROSE
J. FRASER ROBERTS
A. RODGER
W. STEPHENSON
P. E. VERNON

Managing Sub-editor J. W. WHITFIELD

Published by
THE BRITISH PSYCHOLOGICAL SOCIETY
Printed and distributed by
THE ABERDEEN UNIVERSITY PRESS LTD.
6 UPPER KIRKGATE
ABERDEEN

Price 20*s. per part* (*U.S.A.* $3.50) *Subscription* 30*s.* 6*d. per volume* (*U.S.A.* $4.50)

PUBLICATIONS OF THE BRITISH PSYCHOLOGICAL SOCIETY

The British Journal of Statistical Psychology is issued by the British Psychological Society. The subscription price per volume is 30*s.* net for Members of the Society and 30*s.* 6*d.* (post free) for non-members. The subscription price in the U.S.A. is $4.50 (post free). Members of the Society should send their subscriptions to THE SECRETARY, British Psychological Society, Tavistock House South, Tavistock Square, London, W.C.1: non-members to the ABERDEEN UNIVERSITY PRESS, 6, Upper Kirkgate, Aberdeen, Scotland. Until further notice the Journal will be issued in two parts each year, in May and November.

Papers for publication should be sent to Sir CYRIL BURT, 9, Elsworthy Road, London, N.W.3. Authors will receive 25 copies of their articles free; extra copies may be purchased provided the order is given when the proofs are returned. Contributors should indicate whether they wish their reprints to be bound in covers. The printers will state current prices when sending out proofs.

Books for review should be sent to the Editor, advertisements to the Secretary, British Psychological Society.

The Society also issues quarterly *The British Journal of Psychology* and *The British Journal of Medical Psychology:* subscriptions, 60*s.* per annum for either journal, should be sent to the Cambridge University Press, Bentley House, Euston Road, London, N.W.1. The Society publishes, jointly with the Association of Teachers in Colleges and Departments of Education, *The British Journal of Educational Psychology:* for this the subscription, £1 per annum, should be sent to Methuen & Co., Ltd., 36, Essex Street, London, W.C.2. Members of the Society receive the foregoing journals on special terms; enquiries should be addressed to the Secretary, British Psychological Society.

Papers for publication in *The British Journal of Psychology* should be sent to Professor James Drever, Department of Psychology, the University, Edinburgh; those for publication in *The British Journal of Medical Psychology* to Dr. J. D. Sutherland, 2, Beaumont Street, London, W.1.; those for publication in *The British Journal of Educational Psychology* to Professor P. E. Vernon, London University Institute of Education, Malet Street, London, W.C.1.

PREPARATION OF PAPERS

Contributors are asked to send their papers in a form which is ready for submission to the printer: that is to say, the arrangement of the material should follow that observed in previous publications of this *Journal.* Each article should be headed either with a series of numbered headings, in italic, corresponding with the cross-headings of the several sections, or with a brief abstract: (the latter procedure is preferred; an example will be found on p. 29 of this issue). Each section should have a short cross-heading, in capitals; and the whole should conclude with a summary embodying the main conclusions reached. Special attention should be paid to such details as correct spelling, grammar, capitalization, numbering, etc. (particularly in the case of tables and references). In preparing manuscripts and in correcting proofs authors should conform, so far as possible, with the 'Recommendations to Authors' set out in *The Printing of Mathematics* by T. W. Chaundy, P. R. Barrett, and Charles Batey (Oxford University Press, 1954, ch. II, pp. 21-73).

Under the new arrangements for printing, the cost for setting up tables and algebraic formulae which require hand composition is nearly four times as much as the cost of machine composition. Contributors are therefore advised to arrange their material so that it can be printed as economically as possible. Often numerical results and algebraic formulations can, with a little simplification, be readily adapted for purposes of machine composition. Manuscripts involving many tables and displayed formulae have to be submitted first to the printers for an estimation of the cost of the various portions, and may eventually have to be returned, after some delay, to the authors for drastic revision.

Minor contributions (e.g., summaries of theses and the like) should be arranged in the form adopted for printing 'Notes'.

In future authors will receive two complete slip-proofs of their papers, but no page proof. They will be expected to pay the cost of all corrections for which they themselves are responsible.

NOTE TO SUBSCRIBERS

The editors and publishers wish to express their regret for the delay in the publication of the current issue: this has been due to unforeseen difficulties incidental to the change of publisher and printer.

BRITISH JOURNAL OF STATISTICAL PSYCHOLOGY
Vol. VIII Part I May 1955

CONTENTS

		Page
OBITUARY	Sir Godfrey Thomson	1
ROBERTS, J. A., AND DUNSDON M. I.	A Study of the Performance of 2,000 Children on Four Vocabulary Tests	3
GUTTMAN, L.	An Additive Metric from all the Principal Components of a Perfect Scale	17
STUART, A.	The Correlation between Variate Values and Ranks	25
CURETON, E. E.	On the use of Burt's Formula for Estimating Factor Significance	28
BURT, C., COOPER, W. F., AND MARTIN, J. L.	A Psychological Study of Typography . . .	29
NOTE	The International Conference on Factor Analysis : Programme of Papers	58
BOOK REVIEW	Digital Computing Machines	59

Printed in Great Britain at The Aberdeen University Press Ltd., 6 Upper Kirkgate, Aberdeen

BRITISH JOURNAL OF STATISTICAL PSYCHOLOGY
Vol. VIII Part II November 1955

CONTENTS

		Page
GUTTMAN, L.	The Determinacy of Factor Score Matrices with Implications for Five Other Basic Problems of Common-Factor Theory	65
CATTELL, R. B., and CATTELL, A. K. S.	Factor Rotation for Proportional Profiles: Analytical Solution and an Example . . .	83
SIMON, H. A., and Guetzkow, H.	Mechanism Involved in Group Pressures on Deviate-Members	93
BURT, C.	Test Reliability Estimated by Analysis of Variance	103
MAHMOUD, A. F.	Test Reliability in Terms of Factor Theory . .	119
	Indexes	136

Printed in Great Britain at The Aberdeen University Press Ltd., 6 Upper Kirkgate, Aberdeen

OCCUPATIONAL PSYCHOLOGY

Editor : ALEC RODGER

October, 1955 Volume 29, No. 4

CONTENTS

Fifty Years of Psychology	By F. C. BARTLETT
A Neglected Factor in Labour Turnover	By ROBERT H. GUEST
Thirty Years of Psychology in an Industrial Firm	By T. M. HIGHAM
Incentives	By SYLVIA SHIMMIN
An Age-Analysis of Some Agricultural Accidents	By H. F. KING
Book Reviews	
Other Books Received	

Annual Subscription 30 *shillings*

National Institute of Industrial Psychology
14, Welbeck Street, London, W.1

HUMAN RELATIONS

A Quarterly Journal of Studies towards the Integration of the Social Sciences

Contents of VOL. VIII, No. 3

A. K. RICE. The Experimental Reorganization of Non-Automatic Weaving in an Indian Mill.

EDITH BECKER BENNETT. Discussion, Decision, Commitment, and Consensus in "Group Decision".

HAROLD H. KELLEY. Salience of Membership and Resistance to Change of Group-Anchored Attitudes.

CYRIL SOFER. Reactions to Administrative Change : A Study of Staff Relations in Three British Hospitals.

JACK BLOCK and LILLIAN BENNETT. The Assessment of Communication : Perception and Transmission as a Function of the Social Situation.

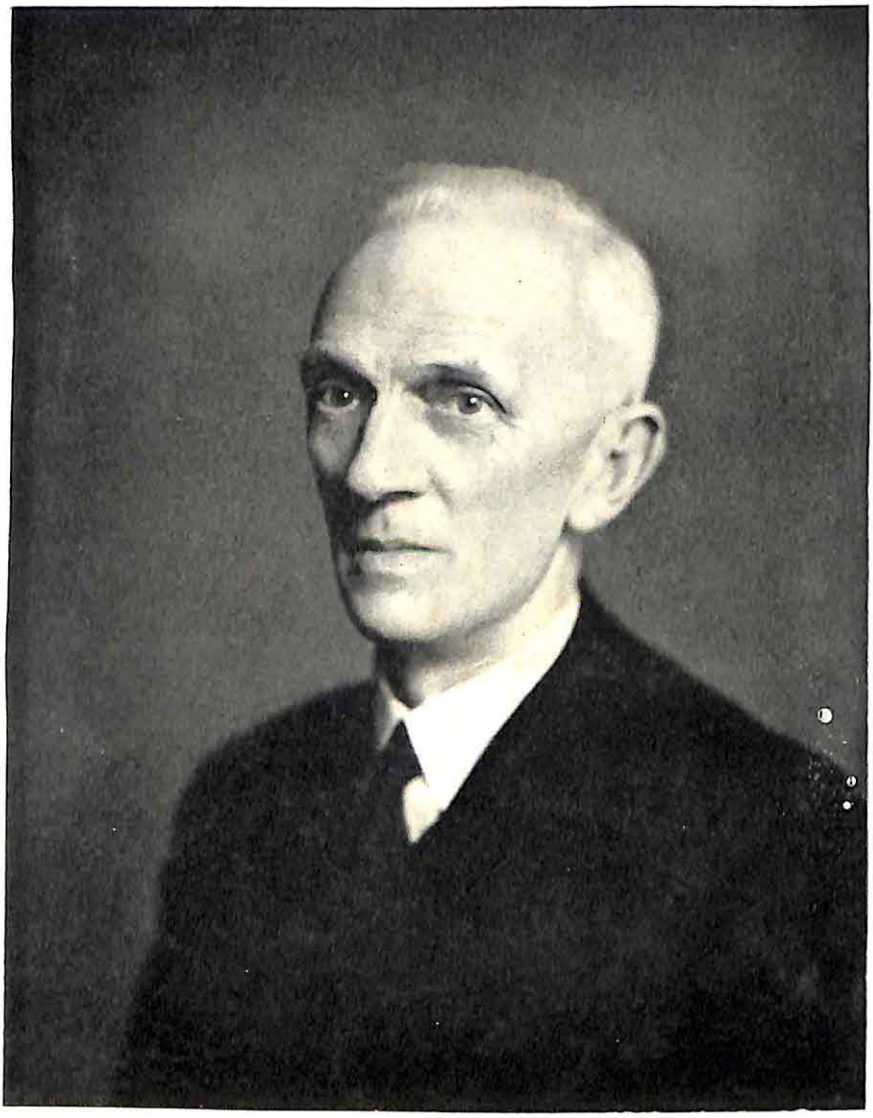
ALBERT PEPITONE and GEORGE REICHLING. Group Cohesiveness and the Expression of Hostility.

Book Reviews.

The Tavistock Institute of Human Relations
London, England.

The Research Center for Group Dynamics
Ann Arbor, Mich.

TAVISTOCK PUBLICATIONS LTD
2 Beaumont Street, London, W.1



SIR GODFREY THOMSON

Vol. VIII
Part I

The British Journal of Statistical Psychology

May
1955

SIR GODFREY THOMSON

Readers will be grieved to learn that Sir Godfrey Thomson, for many years one of the editors of this *Journal*, died in hospital on February 9th. He had long been regarded as 'the most eminent living representative of Scottish education'. Though not himself a Scot, he was born at Carlisle, not far from the Scottish border, on 27 March 1881. For the earlier part of his life his home was on Tyneside. But in 1925 he crossed the border, and the remainder of his years were spent in Edinburgh.

In his autobiography he relates how he 'went to the local Board School, and won a scholarship which provided books and free tuition at Rutherford College in Newcastle-on-Tyne'. At the age of sixteen he returned to his old school as a pupil teacher 'with a salary of £15 for the first year'.[1] Two years later he won a 'Queen's scholarship' which gave pupil teachers free tuition for a degree; and chose the Durham College of Science. In 1903 he obtained the B.Sc. with distinction in Physics and Mathematics. He then took a post as physics master in a Roman Catholic School, but shortly afterwards obtained another scholarship which enabled him to study in Germany. At that time he believed that his own future lay in the field of wireless telegraphy. He therefore decided to go to Strasburg, where the foremost German expert in the subject, Ferdinand Braun, held the Chair of Physics. In 1906 he received the doctorate *summa cum laude* for a thesis *Über den Durchgang Hertzscher Wellen durch Gitter*.

The conditions attached to the Queen's scholarship required holders, after due qualification, to teach for nine years 'in an Elementary School, the Army, the Navy, or the Workhouse'. Accordingly Thomson returned to Armstrong College at Newcastle to take up a post as lecturer to students intending to teach in elementary schools, which was accepted as equivalent to teaching there oneself. Among other courses he now had to deliver lectures on educational psychology. For this he determined to equip himself by spending the summer vacation under C. S. Myers in the Psychological Laboratory at Cambridge, where I myself first had the pleasure of meeting him. He relates how Myers 'one day brought me William Brown's little book, *The Essentials of Mental Measurement*' (to which Thomson himself later contributed several chapters), and how it aroused his interest in psychological methods. As a result he and his wife spent several months collecting data on weight-discrimination; and in 1912 he submitted the results to the Editor of the *British Journal of Psychology*. The manuscript, he tells us, was first returned with a criticism from the referee, saying 'the writer does not seem to know of F. M. Urban's work'. However, with warm encouragement from Myers, who had already decided that "the young man from the North was one of his ablest students", he set himself to study, and eventually to correct Urban's formula, "thus placing the 'constant process' on a basis which made it the precursor of the modern 'probit method'".

He next turned his attention to the second part of Brown's book. From 1904 onwards, under the influence of McDougall at Oxford, Brown, Spearman, and I had been trying to apply Pearson's new method of correlation to measurements obtained from mental tests; but as Brown reports, our conclusions were widely different: whereas I found both general and group factors, Spearman found only a general factor, and Brown found nothing but group factors. Spearman replied to Brown's criticisms in a long paper, written jointly with Hart, in which, as Thomson puts it, he made "sweeping claims for the theory that all correlations between psychological tests were due to only one factor, *g*". Thomson had the happy idea of representing the rival hypotheses by patterns of dice, and was thus able to show that "it was possible to make a set of artificial test scores, without any general factor, which nevertheless gave correlations fully satisfying Spearman's criterion for the Two-Factor theory". The account was published in 1916, and formed the first of a long series of highly original papers on factorial techniques. In his own opinion his most important discovery was his

[1] *A History of Psychology in Autobiography*, Vol. IV (H. S. Langfield, ed., 1952).

proof that "the random interplay of a large number of small independent influences will produce a matrix of correlations which can be reduced to a low rank, or even to a rank of unity"; and on this he based his celebrated 'sampling theory'.

After the first world war was over, Pearson offered him a post in the Galton Laboratory at University College; but having been already assured of a Chair of Education at Newcastle, Thomson eventually decided—to the great disappointment of those of us who were working in London—to decline the offer. During the next six years he was mainly occupied in preparing his books on *Instinct, Intelligence and Character* and *A Modern Philosophy of Education*—both based on his courses as Professor of Education.

Meanwhile, Dr. Andrew Messer, the Chairman of the Education Committee for Northumberland (the county whose capital is Newcastle), persuaded the local authority to introduce a scheme of mental testing similar to that already established in London. Thomson was invited to assist in the project, and thus began the series of 'Northumberland Tests', in which I was occasionally asked to co-operate. In this way I myself became a frequent visitor to his house; and like all who came into touch with him, profited greatly from his keen and stimulating discussions of educational and statistical problems. There are few to whom I owe so much.

In 1925 he was invited to accept the Directorship of the Teacher Training College at Moray House—a post which carried with it the Chair of Education at Edinburgh. Aided by a devoted succession of research-students he continued the compilation of mental tests, now re-named 'Moray House Tests'; and generously handed over all profits to a trust, which was later able to endow a lectureship in experimental education. In his new post, he tells us, his interest in mathematical techniques was revived 'by the presence in Edinburgh University of Edmund Whittaker and A. C. Aitken, who were making fundamental contributions to this subject'—particularly in the field of matrix algebra. And later, a Sabbatical year, together with a grant from the Carnegie Corporation, enabled him to spend twelve months writing his best-known work, *The Factorial Analysis of Human Ability*. In this country it is still by far the most widely studied textbook on the subject.

Throughout this period he gave active help to the Scottish Council for Research in Education. In 1932, under the chairmanship of the Ayrshire Director of Education, the Council organized a mental survey somewhat on the lines of those carried out by Lewis and others in England, but covering a complete age-group; and in 1947, under the chairmanship of Thomson, the survey was repeated on still more ambitious and fruitful lines. The planning of the scheme and the supervision of the subsequent reports occupied most of his leisure time in the years that followed.

During the first half of the century, Britain was fortunate in having a number of scientists, many of international repute, who were interested in the application of statistical methods to psychological problems—some were statisticians, like Pearson, Yule, W. F. Sheppard, and M. S. Bartlett, others psychologists or educationists like Spearman, Nunn, Brown, Maxwell Garnett, and Thomson himself—most of them, alas, no longer with us. Nearly all were members of the British Psychological Society, and were able, just before the second war, to persuade the Council to consider publishing a new periodical dealing specifically with mathematical techniques. The war itself deferred its appearance. But in 1947 publication was started with Thomson as one of the editors. Until his health compelled him to relinquish the work, he proved an invaluable collaborator, severe in the standards which he set before us, tireless in helping junior authors, active in seeking publications from overseas. He himself contributed several papers to the earlier volumes, each a model of terse and lucid presentation. On retiring from his Chair, he planned once again to take up the statistical problems that had so often engaged his attention. His brilliant little book on *The Geometry of Mental Measurement*, published only last year, is an outline of what he had in view. We hope in a later issue to print a fuller review of his many lasting contributions to the field of education and psychology. Here it is only possible to express the grief which teachers, psychologists, educationists, and a large company of personal friends both in this country and abroad will feel at his irreparable loss.

CYRIL BURT

Vol. VIII Part I The British Journal of Statistical Psychology May 1955

A STUDY OF THE PERFORMANCE OF 2,000 CHILDREN ON FOUR VOCABULARY TESTS

I. GROWTH CURVES AND SEX DIFFERENCES

By M. I. DUNSDON and J. A. FRASER ROBERTS
Burden Mental Research Department, Stoke Park Hospital, Bristol

I. *Introduction: Material and Methods.* II. *Results:* (*a*) *Basic Data;* (*b*) *Linearity of the Growth Curves: Increment of Words with CA;* (*c*) *Equations Relating CA and Words Defined;* (*d*) *Sex Differences;* (*e*) *Intercorrelations of the Four Vocabularies;* (*f*) *Comparison with the Previous Sample.* III. *The Inclusion of Independent and Private Schools.* IV. *Absentees.* V. *Summary.*

I. INTRODUCTION: MATERIAL AND METHODS

In a previous paper (Dunsdon and Roberts, 1953) reasons were given why it was desired to use vocabulary scales as quickly-given individual tests, not for the assessment of single children, but for the comparison of groups. For the comparison of large groups vocabulary tests have manifest advantages in respect of reliability, validity, relative absence of practice effect, and, above all, in the amount of information obtained per unit of testing time. In order to study in some detail the performance of children on vocabulary tests, and to establish norms for comparison, it was decided to apply four scales to a relatively large sample, which it was also hoped would be effectively random. By using four vocabularies the consistency of the results can be assessed, alternatives are made available, or, on the other hand, two or more vocabularies can be combined.

The four oral definition vocabularies used were respectively that of the Terman Merrill (1937 Revision) Form L, Raven's Mill Hill Vocabulary Tests A and B, and that from Wechsler's Intelligence Scale for Children. We are greatly indebted to Professor L. M. Terman, Mr. J. C. Raven and Dr. D. Wechsler for permission to use their vocabularies for this piece of work. In the subsequent text, the four vocabularies will be referred to as TML, MHA, MHB and WISC.

Mention should be made of the way in which the vocabularies were administered and scored. In the case of TML and WISC the usual procedures were employed; but, after the usually accepted limit of 5 successive failures, further words were read to the child, who was instructed to nod or otherwise indicate if he thought he could attempt a definition. The necessity for such a modification of testing procedure was referred to by Kennedy-Fraser in the report on the use of the Terman-Merrill Intelligence Scale in Scotland (1945). With Mill Hill A and B, both tests were given in the oral definition form only, and selection of synonyms was not used. As regards the limiting point, the same procedure was employed as with TML and WISC.

For all four vocabularies the scoring was 1 point for each word satisfactorily defined. With the WISC vocabulary this involved a modification of the usual procedure, in which half points may be credited for relatively poorer definitions. The bulk of the testing was carried out by one of us (M. I. D.), the remainder by Miss M. E. Drabble and Mr. J. Annett.

The plan was to visit all the schools in the City of Bristol, together with some others outside the city boundary at which it was expected that Bristol children might be found. The children tested were those between the ages (in completed months) of 5 years 0 months and 14 years 11 months inclusive, whose birthdays fell on the first day of any calendar month. This provides a sample of approximately 3 per cent. Children whose homes lay outside the boundaries of the city were excluded. It was necessary to limit the survey to the ages at which school attendance is compulsory, for outside these limits it is practically impossible to avoid incomplete and almost inevitably biased samples.

It was found during the course of the Scottish Survey (Scottish Council for Research in Education, 1953) that sampling by day of birth in the month is highly efficient for obtaining a random sample. There was no significant difference between the total sample and the subsample in regard to 5 variables. One important advantage of this method of sampling over sampling by date of birth, appears when the testing occupies a considerable period, in this instance nearly two years. If sampling is by date of birth, the children are growing throughout the period of testing; the number of schools is finite, hence considerable bias may be introduced by the order in which they are visited.

It was found that 10 per cent. of the boys and 7 per cent. of the girls were at private and independent schools; and it is shown later in this paper how greatly the results would have been affected had these children been omitted. Another possible source of bias is failure to include absentees. Accordingly further visits were made to trace and test children who had been absent when the school was first visited. Some data on the effect of this are also given later.

Thanks to the generous co-operation of the authorities concerned and of the Headteachers, every school in the administrative area was visited, including those of the Education Authority and the independent and private schools. It is perhaps possible that one or two very small private schools may have been overlooked, though we do not think so. The small proportion of mentally defective children and children excluded from school was also included. Inevitably there must be some loss of children attending schools outside the boundaries of the city. This was minimized as far as possible by searching for Bristol children at Education Authority schools outside the area, as well as at a number of independent schools, both residential and non-residential.

The method of sampling is flexible. The limits of age and the place of residence are fixed in relation to the date on which the school is visited. During the period of testing children are moving into and out of the city. They are increasing in age, so that CA depends upon the date of the visit; children are also entering and leaving the age-span of the survey. All this is automatically allowed for, the sample representing the average for the city over a period of time and not at a given moment. The pupils are represented in proportion to their numbers in the population; and the unequal numbers at different ages seen in Table I clearly reflect fluctuations in the birthrate between 1938 and 1949.

There are two difficulties, however. The first is that children do not always start school immediately after their fifth birthday; hence the youngest age-group, 5·0 to 5·5, is not completely represented. Perhaps we should have omitted this age-group. We were anxious, however, to establish norms down to 5 years, and we

TABLE I. BASIC DATA

A. Numbers: Boys, 980; Girls, 967

B. Sums: CA (in months) and Words

	Boys	Girls
CA	112,550	110,996
TML	11,747	10,791
MHA	10,451	9,743
MHB	10,115	9,249
WISC	15,638	14,255

C. Sums of Squares and Cross Products: CA (in months) and Words [1]

Boys	CA	TML	MHA	MHB	WISC
CA	14,031,414	1,500,497	1,345,985	1,309,351	1,928,554
TML		175,675	155,567	152,242	215,952
MHA			141,513	137,042	193,401
MHB				135,555	188,674
WISC					276,614

Girls	CA	TML	MHA	MHB	WISC
CA	13,865,678	1,381,831	1,265,106	1,208,350	1,762,325
TML		149,073	134,603	129,733	182,482
MHA			125,531	119,549	166,915
MHB				116,769	160,138
WISC					232,961

[1] Corresponding figures for combinations of more than one vocabulary can be obtained algebraically.

D. Sums of Words by 6 month Groups of CA

CA		Boys					Girls				
Yr.	Mth.	NO.	TML	MHA	MHB	WISC	NO.	TML	MHA	MHB	WISC
5	0- 5	43	233	154	138	396	38	179	118	101	308
	6-11	55	319	236	204	541	80	454	285	242	733
6	0- 5	54	402	296	283	627	58	360	255	250	578
	6-11	69	481	386	374	783	56	393	321	298	621
7	0- 5	50	395	326	313	623	61	458	369	344	690
	6-11	62	522	476	457	794	51	394	358	327	603
8	0- 5	61	551	495	486	830	49	440	395	391	649
	6-11	60	663	618	578	954	41	402	371	350	575
9	0- 5	57	621	581	544	895	53	532	543	503	802
	6-11	49	568	554	520	808	57	609	608	548	865
10	0- 5	55	758	690	673	989	55	687	658	609	915
	6-11	49	668	630	618	900	49	651	625	588	831
11	0- 5	40	525	497	468	684	44	567	560	512	755
	6-11	39	613	558	535	753	33	494	470	447	595
12	0- 5	37	617	570	571	743	44	677	596	592	802
	6-11	37	640	573	555	741	42	721	666	642	823
13	0- 5	43	780	684	696	908	44	731	682	669	844
	6-11	37	745	647	632	814	31	521	475	468	598
14	0- 5	43	819	719	719	940	38	709	639	631	779
	6-11	40	827	761	751	915	43	812	749	737	889

trust that no appreciable bias has been introduced. The second difficulty is more serious. This is the question of absentees on the day of the visit. The absentees falling within the sample were duly recorded and visits were made later—a troublesome and time-consuming task from the practical point of view; with some persistent absentees as many as six or seven visits were required. There is also the question of bias being introduced by the loss of absentees. Actually, despite assiduous pursuit, 27 boys and 45 girls in the older age-groups left school before they could be tested. At the other end of the range, absentees just over 5 years were tested when they had grown rather older, and, of course, were not replaced by corresponding children who were less than 5 at the time of the original visit. We trust, however, that these losses have not had an appreciable effect. In fact, as is shown later, in our sample the difference in performance between absentees and the remainder has proved surprisingly small.

Even though, however, the loss of some absentees (a loss which is not independent of CA) may not have made any appreciable difference to the final results, an improvement in experimental design seems called for should a future survey be planned on the same lines. It might be possible, during the original visits to schools, to define samples composed of children present that day, but absent on some previous date or dates arbitrarily selected. In this way a sample of absentees could be built up and the main sample supplemented by appropriate addition of figures from the absentee sample. If it proved practicable, this plan would ensure completeness, as well as save a great deal of time.

There is one other small loss. Twenty children, whose names were on school registers, could not be traced. They were drawn, in the main, from families apt to make sudden and spasmodic departures from the area, without notifying the school authorities of their removal or their return.

To sum up, the losses were relatively small in relation to the size of the sample; and, furthermore, while it cannot be said that they are unbiased, it may be hoped, for the reasons given, that there has been no appreciable effect on the results set out in this paper, and that the 2,000 children do represent a nearly random sample of the schoolchildren of a large city.

This paper deals with the first analysis; the question of norms and a more detailed examination of the scores will follow later.

II. RESULTS

(*a*) *Basic Data.* Basic data for the complete samples of boys and girls respectively are given in Table I. The figures given are those required for the analyses made in this paper. The actual bivariate frequency distributions will be more appropriately included, to the extent that space allows, in the second paper of this series, which will deal with the establishment of norms.

(*b*) *The Linearity of the Growth Curves: Increment of Words with CA.* It is shown in the next section that there are substantial differences in performance between boys and girls, hence they are treated separately here. For the present purpose 6-month groups of CA are used, as shown in Table I D. The fitting is shown in Table II.

It turns out that with TML and MHB there is no significant departure from linearity of increment of mean score in either sex. This is also true of MHA, though for boys only. With girls the departure from linearity in MHA somewhat exceeds the 1 per cent. level of significance. With WISC, the growth curve is significantly non-linear for both sexes, the variance ratios almost attaining the 1 per cent. point.

The main feature of the departure from linearity with WISC, as will be seen from Fig. 1, is a tendency for the rate of increment to fall off with advancing CA. In view of the similarity of the results in boys and girls, it may indeed be that this is a feature of the WISC; and in establishing norms for this vocabulary it may prove inadvisable to rely on the linear regressions. The result for MHA in girls is puzzling. There is certainly no obvious reason why the sexes should differ in this respect; and in spite of the high significance of the departure, we are rather inclined to think that it may perhaps be due to chance. It should also be noted that when MHA and MHB are used with oral definition, they are intended to be combined into a single test: a point we shall not overlook at a later stage when considering norms. It is, however, striking that, when all four vocabularies are added together, the departures from linearity are much below the 5 per cent. level of significance. Had there been a general tendency towards the same kind of curvilinearity, although non-significant in some vocabularies, it should nevertheless have been revealed when they were combined. In oral definition tests, therefore, it seems unlikely that curvilinearity is a characteristic feature of increment with age.

When vocabulary tests were first introduced, they were intended to estimate total vocabulary, the choice of words being made, for example, at random from a dictionary. It would be interesting to find, if indeed it is true, that increase in total vocabulary is, on the average, a linear function of age.

(*c*) *Equations Relating CA and Words Defined.* From this point onwards we have used single months as units of CA. The equations derived from the data of Table I are as follows:

TML	Boys	$y = -\ 3{\cdot}742 + 0{\cdot}136956x,$
„	Girls	$y = -\ 3{\cdot}450 + 0{\cdot}127273x,$
MHA	Boys	$y = -\ 4{\cdot}476 + 0{\cdot}131826x,$
„	Girls	$y = -\ 4{\cdot}897 + 0{\cdot}130445x,$
MHB	Boys	$y = -\ 5{\cdot}021 + 0{\cdot}133595x,$
„	Girls	$y = -\ 5{\cdot}403 + 0{\cdot}130398x,$
WISC	Boys	$y = +\ 2{\cdot}183 + 0{\cdot}119937x,$
„	Girls	$y = +\ 1{\cdot}879 + 0{\cdot}112059x,$
Combined vocabularies	Boys	$y = -\ 11{\cdot}057 + 0{\cdot}522314x,$
„	„ Girls	$y = -\ 11{\cdot}871 + 0{\cdot}500174x,$

where y is number of words and x is CA in completed months.

(*d*) *Sex Differences.* Sex differences appear consistently in the results from all four vocabularies. These are illustrated in Figs. 1 and 2, which show the observed means for complete years of CA, together with the fitted linear regressions.

The difference in mean CA between boys and girls is very small, in fact only ·06 months. The regressions obtained by taking single months of CA as in the preceding section, have nevertheless been used to adjust the means of both sexes to age 114·8 months. The comparison is shown in Table III. It will be seen that the means at fixed age for boys exceed those for girls in all four vocabularies. The differences, ranging from 0·6 to 1·2 words at about 9½ years, may not appear very large in terms of absolute words; nevertheless all are very highly significant.

Turning to the regressions, the differences, as shown in Table IV, are much smaller. Boys show a more rapid rate of growth with all four vocabularies; but with all except TML the difference is not significant: with TML it just attains the 5 per cent. level. The difference in total increment is non-significant with each of the four vocabularies. Thus, the performance of boys exceeds that of girls at all

TABLE IIA. TESTS FOR LINEARITY OF INCREMENT OF SCORE WITH CA (6 MONTH GROUPS OF CA)

Variation of Vocabulary Score	Degrees of Freedom	Sum of Squares (words)^2	Mean Square (words)^2	Variance Ratio	Significance of Deviation
TML Boys					
Linear regression	1	20,710·3			
Deviations from L.R.	18	393·4	21·86	1·52	—
Within arrays	960	13,763·1	14·34		
Total	979	34,866·8			
TML Girls					
Linear regression	1	18,184·8			
Deviations from L.R.	18	218·1	12·12	1·12	—
Within arrays	947	10,250·6	10·82		
Total	966	28,653·5			
MHA Boys					
Linear regression	1	19,213·4			
Deviations from L.R.	18	272·7	16·26	1·48	—
Within arrays	960	10,554·4	10·99		
Total	979	30,060·5			
MHA Girls					
Linear regression	1	19,136·5			
Deviations from L.R.	18	316·1	17·56	2·10	+ +
Within arrays	947	7,912·9	8·36		
Total	966	27,365·5			
MHB Boys					
Linear regression	1	19,693·2			
Deviations from L.R.	18	257·8	14·32	1·23	—
Within arrays	960	11,202·7	11·67		
Total	979	31,153·7			
MHB Girls					
Linear regression	1	19,134·6			
Deviations from L.R.	18	190·0	10·56	1·11	—
Within arrays	947	8,981·1	9·48		
Total	966	28,305·7			
WISC Boys					
Linear regression	1	15,858·3			
Deviations from L.R.	18	378·6	21·03	1·86	+
Within arrays	960	10,839·3	11·29		
Total	979	27,076·2			
WISC Girls					
Linear regression	1	14,086·3			
Deviations from L.R.	18	3,08·1	17·12	1·92	+
Within arrays	947	8427·0	8·90		
Total	966	22,821·4			

TABLE IIB. TESTS FOR LINEARITY OF INCREMENT OF SCORE WITH CA (6 MONTH GROUPS OF CA): TOTALS FOR BOYS AND GIRLS

Variation of Vocabulary Score	Degrees of Freedom	Sum of Squares (words)^2	Mean Square (words)^2	Variance Ratio	Significance of Deviation
TOTAL. All Vocabularies. Boys.					
Linear regression	1	301,166			
Deviations from L.R.	18	4,089	227·2	1·33	—
Within arrays	960	163,635	170·5		
Total	979	468,890			
TOTAL. All Vocabularies. Girls.					
Linear regression	1	281,111			
Deviations from L.R.	18	2,779	154·4	1·20	—
Within arrays	947	121,756	128·6		
Total	966	405,646			

the ages studied; the difference, however, does not widen greatly with advancing CA.

It seems clear, therefore, that tests involving oral definition as such definitely favour boys; it is not merely a question of a particular vocabulary being sex-biased.

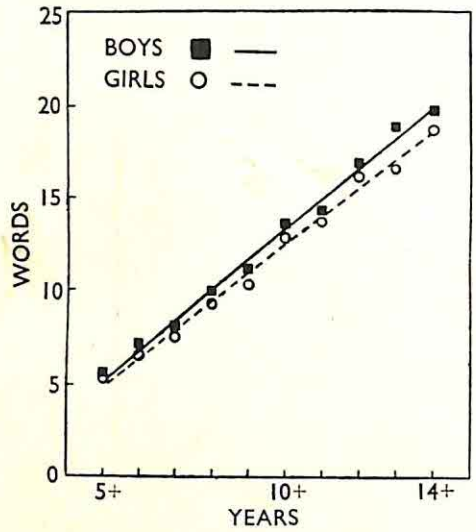
TABLE III. MEAN NUMBER OF WORDS DEFINED AT FIXED CA (114·8 MONTHS)

Vocabulary	Boys	Girls	Difference Boys — Girls	S.E. of Diff.	Diff./S.E.
TML	11·980	11·161	+ 0·819	0·1610	5·09
MHA	10·658	10·078	+ 0·580	0·1419	4·09
MHB	10·315	9·567	+ 0·748	0·1475	5·07
WISC	15·952	14·743	+ 1·208	0·1448	8·34
Total	48·905	45·549	+ 3·356	0·5547	6·05

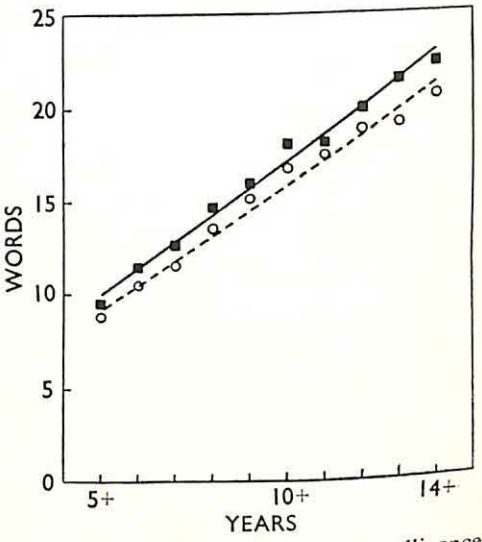
TABLE IV. INCREMENT IN WORDS PER YEAR OF CA

Vocabulary	Boys	Girls	Difference Boys — Girls	S.E. of Diff.	Diff./S.E.
TML	1·643	1·527	+ 0·116	0·0571	2·03
MHA	1·582	1·565	+ 0·017	0·0504	0·34
MHB	1·603	1·565	+ 0·038	0·0523	0·73
WISC	1·439	1·345	+ 0·094	0·0514	1·83
Total	6·267	6·002	+ 0·265	0·1969	1·35

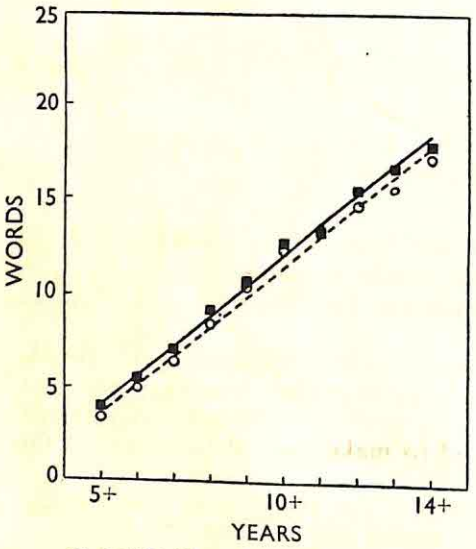
A personal communication from Mr. Raven is very pertinent at this point. He tells us that, in making up MHA and MHB, he was careful to reject any words



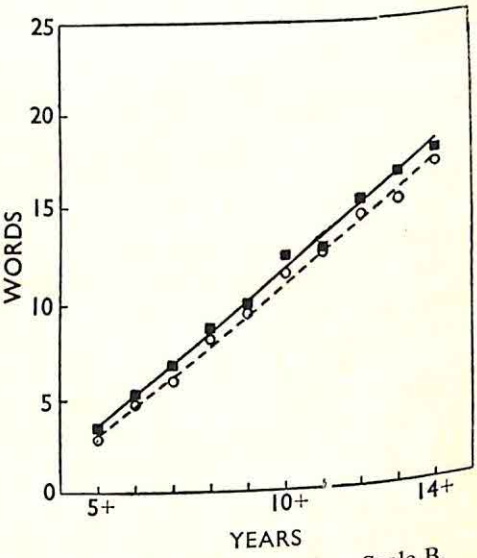
A. Vocabulary form Terman-Merrill Scale, Form L.



B. Vocabulary from Wechsler Intelligence Scale for Children.



C. Mill Hill Vocabulary Scale A.



D. Mill Hill Vocabulary Scale B.

FIG. 1.—Mean Numbers of Words Defined by Years of CA, with Fitted Linear Regressions.

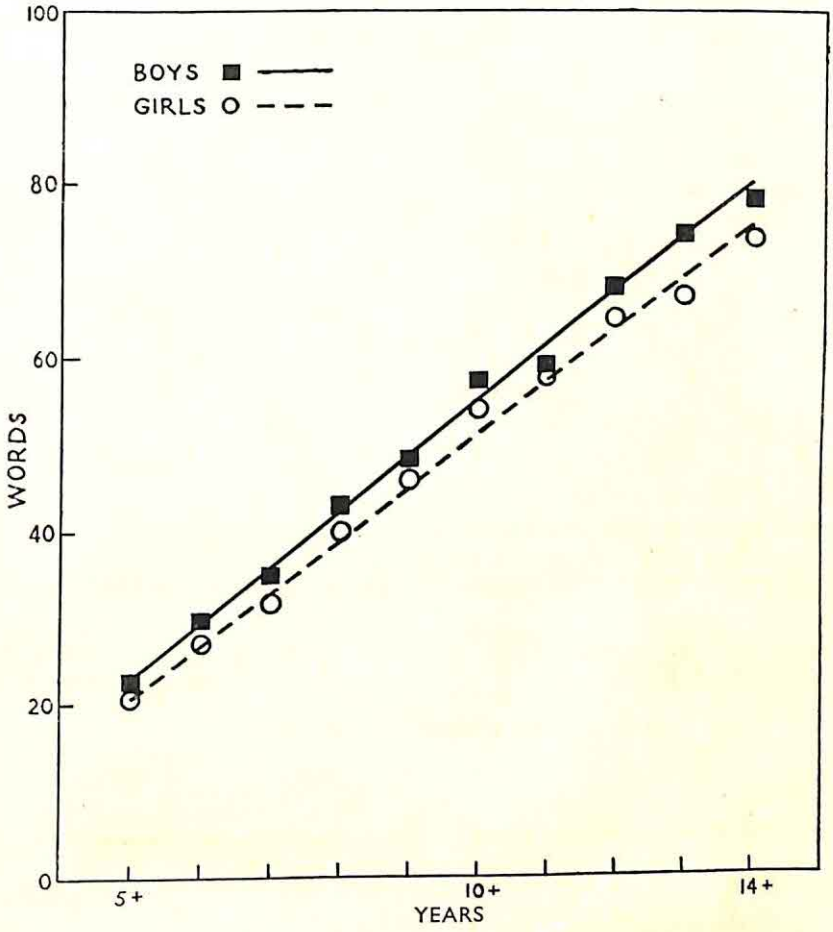


FIG. 2.—All four Vocabularies combined. Mean Number of Words Defined by Years of CA, with Fitted Linear Regressions.

which he suspected might be even slightly more familiar to one or other sex. Nevertheless, the sex difference in his vocabularies is much the same as that shown by TML and WISC, in which, as Professor Terman and Dr. Wechsler inform us, no such exclusion of words was practised.

Another sex difference, of a more familiar kind, also appears in our sample. It has often been noted that boys are more variable in performance than are girls; and this emerges in a striking way from the present data. The detailed variances at fixed CA are shown in Table V. It is hoped to make more detailed use of this information in another paper.

(*e*) *Intercorrelations of the Four Vocabularies.* The intercorrelations between the scores on the four vocabularies at fixed CA are shown in Table VI.

The intercorrelations seem reasonably high, particularly when it is recalled that they refer to tests taking only a few minutes to give. Naturally they are lower than reliability coefficients based on the repetition of the same vocabulary; these, as mentioned in the previous paper, seem to be generally about ·9. In a sense, the

intercorrelations may be regarded as another kind of reliability coefficient, measuring the consistency of performance in the task of word definition generally, apart from consistency in regard to a particular set of words.

The intercorrelations are appreciably higher with the boys. No doubt this is a reflection of the wider dispersion of their performance, as shown in Table V.

TABLE V. VARIANCE OF WORDS DEFINED AT FIXED CA

	Boys	Girls	Ratio Boys/Girls
TML . .	14·45	10·81	1·34
MHA . .	11·09	8·52	1·30
MHB . .	11·68	9·51	1·23
WISC . .	11·43	9·01	1·27
Total . .	171·09	128·67	1·33

For both boys and girls the intercorrelations show relatively little variation in magnitude, except that, as might be anticipated, the correlation between the two forms of the Mill Hill test (MHA and MHB) gives the highest figures.

TABLE VI. INTERCORRELATIONS OF THE SCORES ON THE FOUR VOCABULARIES

Partial Correlations at fixed CA

Boys	MHA	MHB	WISC
TML . .	·8347	·8476	·8233
MHA . .		·8717	·8314
MHB . .			·8457
Girls	**MHA**	**MHB**	**WISC**
TML . .	·7775	·8024	·7730
MHA . .		·8317	·8094
MHB . .			·8234

(*f*) *Comparison with the Previous Sample.* The present results can be compared, for TML only, with those obtained from the previous sample (Dunsdon and Roberts, 1953). For the means, adjusting to 10·80 years, as used with the previous sample, the figures are as follows:

	Present sample	*Previous sample*
Boys	14·01	13·43
Girls	13·04	12·81

The figures for increment of words per year of CA are:

	Present sample	*Previous sample*
Boys	1·64	1·67
Girls	1·53	1·55

The regressions are thus almost identical; but the mean performance of the present sample of children is ·58 ± ·28 words higher in boys and ·23 ± ·27 words higher in girls. In view of the very different ways in which the samples were made up, it is satisfactory to find that the discrepancies are so small. The sample of 500 test-results used in the previous paper was made up from available report-forms for children who had been tested for a wide variety of purposes, these being selected (using random processes) so as to give a group with a mean I.Q. of 103 and a standard deviation of 17. A large number of the forms, however, were those of brothers and sisters of mental defectives, and few independent school children were included. It might well have been anticipated that a highly verbal test, such as vocabulary, would reflect the lower social background, which such a sample might be expected to show, more clearly than would the Terman-Merrill Scale as a whole. It is, therefore, reassuring to find that the difference between the results in the earlier sample and those obtained from an effectively random sample is barely significant in the boys and not significant in the girls. The sex difference pointed out in the previous paper is now, of course, amply confirmed.

It may also be noted that the correlation between TML vocabulary score and MA on the whole scale, at fixed age, as then discovered, was ·84—a figure which is of the same order of magnitude as the correlation of the TML vocabulary score with the three other oral definition tests, as shown in Table VI.

III. THE INCLUSION OF INDEPENDENT AND PRIVATE SCHOOLS

In our sample 10 per cent. of the boys and 7 per cent. of the girls attended independent and private schools. The results for these children separately are not of much general interest. In other cities the proportions and kinds of children attending these schools might well be quite different. Moreover, there are features in the selection from these schools which make the results somewhat unreal. For example, mean age is substantially higher than in the remainder—19 months in the boys and 14 months in the girls. This reflects a tendency for children to enter these schools at later ages—for example, the very bright children awarded special places as the result of the examination at 11 +. The regressions therefore are much distorted. What is important is to see what the effect would be if a survey of the present kind were to be restricted to children attending the Local Education Authorities schools. It is sufficient to quote the results for one vocabulary only, namely, TML. For boys, the figures are as follows:

	Whole Sample	*Sample omitting Independent and Private Schools*
Mean score at 114·8 months . .	11·98	11·37
Increment of words per year . .	1·64	1·46
Variance of words at fixed CA . .	14·45	10·14

The differences are rather less for the girls, but still substantial. It is clear, therefore, that the omission of independent and private schools would very seriously alter the results, and lead to norms that would be appreciably different from those of the whole child population.

IV. ABSENTEES

The results for absentees (74 boys, 76 girls) are in sharp contrast to those just described. Absentees are lower in performance; but the differences are not great, and omission would in fact have made relatively little difference. The results, again for TML, are as follows:

Boys	*Whole Sample*	*Sample omitting Absentees*
Mean words at 114·8 months . .	11·98	11·99
Increment of words per year . .	1·64	1·66
Variance at fixed CA . . .	14·45	14·61
Girls		
Mean words at 114·8 months . .	11·16	11·22
Increment of words per year . .	1·53	1·55
Variance at fixed CA . . .	10·81	10·97

A number of earlier studies have shown that children absent on the date of school visiting are often appreciably lower in score than the remainder; and we were surprised to find relatively modest differences in the present sample. It may be that the pattern of absenteeism is changing. To mention two possible reasons for this, there are the material benefits now offered by the schools (for example, free meals), and there is the fact that more mothers are out at work during the day. Nevertheless, one would hesitate to suggest that absentees could safely be omitted from a survey of this kind—at least until the relative lack of difference had been confirmed with other samples in other areas. It may be that our sample of absentees proved rather peculiar in this respect; and, until a good deal of confirmatory evidence is forthcoming, there would always be a doubt whether the results had not been fairly seriously distorted. Moreover, even in our sample, the effect of omission of absentees, though relatively small, cannot, in the girls at least, be called negligible.

V. SUMMARY

1. A 3 per cent. sample of the schoolchildren of the City and County of Bristol was selected, by visiting all schools (including a number outside the city) and choosing those children, between the ages of 5·0 to 14·11 years, whose homes were within the city and whose birthdays fell on the first day of any calendar month. Four vocabulary scales were used: that from the Terman-Merrill Scale, Form L; Mill Hill Vocabularies A and B; and that from the Wechsler Intelligence Scale for Children.

2. The present paper gives the first results. The establishment of norms will be dealt with later.

3. All four vocabularies proved to be sex-biased in favour of boys, who on the average defined more words at every age studied. The differences were fairly substantial; and all were highly significant.

4. In 5 out of 8 comparisons given by 2 sexes and 4 vocabularies the increment of score with CA showed no significant departure from linearity. Three showed some divergence. There was no significant departure from linearity in the combined score for all vocabularies.

5. The intercorrelations of the four scores, reduced to fixed age, were of the order of rather more than ·8; they showed little variation in magnitude.

6. It is shown that omission of children attending private and independent schools would have distorted the results very seriously: on the other hand, absentees on the dates of the original visits to the school were not greatly lower in performance than the remainder, and their omission, in this particular sample, would not have altered the results very appreciably. It is not suggested, however, that this would be in general a safe procedure.

Acknowledgments. We are most grateful to Professor Lewis M. Terman and to the Houghton Mifflin Company for permission to use the Vocabulary test from the Terman-Merrill Intelligence Scale (1937); to Mr. J. C. Raven for permission to use the Oral Definitions Form of his Mill Hill Vocabulary Scale; and to Dr. David Wechsler for allowing us to use the Vocabulary from his Intelligence Scale for Children. Special thanks are due to the Bristol Education Committee and to the Chief Education Officer, Mr. G. H. Sylvester, who gave permission for work to be carried out in their schools, and in particular to the Headmasters and Headmistresses of both the Local Education Authority's schools and Independent Schools, without whose willing and active co-operation the children would not have been available.

In the collection of data and in its subsequent analysis, helpful assistance has been given by Miss M. E. Drabble and Mr. J. Annett, Assistant Psychologists at the Burden Mental Research Department. We are indebted to Miss L. D. Buswell for drawing the figures.

REFERENCES

1. Dunsdon, M. I. and Roberts, J. A. F. (1953). 'The Relation of the Terman-Merrill Vocabulary Test to Mental Age in a Sample of English Children.' *Brit. J. Stat. Psychol.*, VI, 61-70.
2. Kennedy-Fraser, D. (1945). *The Terman-Merrill Intelligence Scale in Scotland.* (Publications of the Scottish Council for Research in Education, No. 23.) London: University of London Press.
3. Scottish Council for Research in Education (1953). *Social Implications of the* 1947 *Scottish Mental Survey.* London: University of London Press.

Vol. VIII
Part I

The British Journal of Statistical Psychology

May
1955

AN ADDITIVE METRIC FROM ALL THE PRINCIPAL COMPONENTS OF A PERFECT SCALE[1]

By LOUIS GUTTMAN
Israel Institute of Applied Social Research

Abstract. A new type of metric is established for perfect scales that is different from the three previously known types of metrics. Given a perfect scale of m types of dichotomous items, each rank of persons is regarded as a point in an m-dimensional space defined by the m non-constant principal components. A non-Euclidean distance function is defined for this space. It is proved that the resulting metric is additive : the non-Euclidean distance between any two ranks i and k is the sum of the distances from i to j and from j to k whenever $i \leq j \leq k$. Treating the principal component space as non-Euclidean may also be useful for the study of non-scale structures and of quasi-scales.

I. *Introduction.* II. *A Numerical Example of Principal Component Scores.* III. *Making each Principal Component's Variance Proportional to its* η^2. IV. *The Non-Euclidean Distance Function and the New Additive Metric.* V. *Possible Applications for Non-Scalable Universes and for Quasi-Scales.* VI. *Proof of the Additive Property of the New Metric.*

I. INTRODUCTION

Three different meaningful metrics are known for a perfect scale, and have been described elsewhere [2, 3]. They are, respectively: (*a*) the basic rank order of the population ; (*b*) the first principal component scores, which maintain the basic rank order, but stretch and contract differences in rank to give a best metric in a certain sense of least squares (as defined in [1]); (*c*) 'absolute psychological types,' or the *intervals* along the basic rank order, defined by the bending points of the successive principal component curves.

In the present paper we shall prove that every perfect scale always has also a fourth type of metric, and that this is an *additive* metric. The new metric is computed in an exact algebraic fashion from *all the principal components* of the perfect scale. In this new metric, if i, j, and k are any three scale ranks, and if j is between i and k, then the distance from i to k is the sum of the distances from i to j and from j to k.

We shall first illustrate the new metric and its additive property by a numerical example, and shall give it an interpretation. Then we shall prove algebraically that the additive property holds for all perfect scales.

II. A NUMERICAL EXAMPLE OF PRINCIPAL COMPONENT SCORES

For illustrative purposes, let us take the example published in [2] of a perfect scale of five types of dichotomous items, where all the principal components have been worked out numerically. Five types of dichotomous items yield six ranks (types) for persons and five

[1] This research was facilitated by an uncommitted grant-in-aid to the writer from the Behavioral Sciences Division of the Ford Foundation.

principal component scores for each rank, apart from a constant score (on the constant principal component). Given that the types of items are all equally frequent in the universe of items, and that the types of persons are all equally frequent in the population of individuals (respondents), the principal component scores of the scale turn out to be numerically as in Table I.

TABLE I. RAW PRINCIPAL COMPONENT SCORES FOR EACH RANK OF PERSONS [1]

Rank of Persons	Principal Component				
	I	II	III	IV	V
5	5	5	5	1	1
4	3	−1	−7	−3	−5
3	1	−4	−4	2	10
2	−1	−4	4	2	−10
1	−3	−1	7	−3	5
0	−5	5	−5	1	−1
η^2	·6	·2	·1	·06	·04
Raw sum of squares	70	84	180	28	252

[1] From [2], Table IV, p. 322.

We have copied the principal components' scores in Table I in the same raw form as given in [2, p. 322]. These are raw scores in the following sense: each principal component is a latent vector of a certain matrix. Latent vectors are defined only up to a constant of proportionality, that is, multiplying any column of scores in Table I through by any (non-zero) constant will yield new scores on what is actually the same principal component; geometrically, the *direction* of the latent vector remains unchanged—except possibly for sign—and only its *length* is affected. The length of any given latent vector in Table I is the square root of the corresponding sum of squares shown in the last row of the table.

III. MAKING EACH PRINCIPAL COMPONENT'S VARIANCE PROPORTIONAL TO ITS η^2

For purposes of exposition in [2] the raw scores in Table I were deliberately computed to be in exact whole numbers. This has made the raw sums of squares—and hence the lengths of the corresponding latent vectors—arbitrarily unequal. For some theoretical purposes, it is useful to make the lengths of all latent vectors equal, say equal to 1; this can be done for the data of Table I by dividing each raw score in a given column by the square root of its raw sum of squares. For each column the sum of squares of the new scores will then always be 1.

For our present purpose, as a first step toward our new metric, we need to modify the lengths of the latent vectors in a different manner, namely, to make each new sum of squares equal to the corresponding η^2. One motive for this is as follows. The components were arrived at in the first place from a maximization problem: given the responses of a population of individuals to a universe of items, find that set of scores—or metric—for the individuals which will maximize the correlation ratio over items [1]. The maximizing scores turn out to be the *first* principal component of the set of responses, and so give the *best metric* in this sense of maximization. But the same equations show that there also exist *next best* down to *worst* principal components, as ranked by the correlation ratios they yield. The sum of squares of all correlation ratios is 1 [cf. 2, p. 322], showing that the

total variance of the responses can be accounted for exactly and in a *linear fashion* by all the principal components. (The linear equations for reproducing the original item responses exactly from all the principal component scores are illustrated in [2, pp. 331 f.].) In this sense (and since the principal components for a perfect scale are always orthogonal to each other), the η^2 associated with a given principal component expresses its proportional contribution to the total variance of the scale.

Therefore, in making the length of each principal component vector equal to the η of that component, we are expressing each component in a way which reflects its importance for accounting for the observed data in a linear fashion. The first component is given the largest variance, the last component the smallest variance, and intermediate components variances of intermediate size.

To convert the raw scores of Table I into the proportional form that we desire, it is necessary simply to divide each row by the square root of its raw sum of squares and then multiply by η. This can be done in one step by multiplying each row by the quantity $\sqrt{\{\eta^2/(\text{raw sum of squares})\}}$. Thus, each raw score in the first column of Table I is to be multiplied by $\sqrt{\{\cdot 6/70\}} = \cdot 09258$. For the remaining columns the multipliers are $\sqrt{\{\cdot 2/84\}}$, $\sqrt{\{\cdot 1/180\}}$, $\sqrt{\{\cdot 06/28\}}$, and $\sqrt{\{\cdot 04/252\}}$ respectively. The new scores are shown in Table II.

Here we are omitting the constant component throughout, because it does not affect our new metric. The correlation ratio is actually zero for the constant component, although its latent root is 1; for each of the other principal components, η^2 is always equal to the corresponding latent root.

TABLE II. PRINCIPAL COMPONENT SCORES[1]

Expressed with Sum of Squares equal to η^2

Rank of Persons	Principal Component				
	I	II	III	IV	V
5	·4629	·2440	·1179	·0463	·0126
4	·2771	−·0488	−·1650	−·1389	−·0630
3	·0926	−·1952	−·0943	·0926	·1260
2	−·0926	−·1952	·0943	·0926	−·1260
1	−·2771	−·0488	·1650	−·1389	·0630
0	−·4629	·2440	−·1179	·0463	−·0126
η^2	·6	·2	·1	·06	·04
Sum of squares	·6	·2	·1	·06	·04

[1] Computed from Table I by multiplying each column of raw scores of Table I by $\sqrt{\{\eta^2/(\text{raw sum of squares of that column})\}}$.

IV. THE NON-EUCLIDEAN DISTANCE FUNCTION AND THE NEW ADDITIVE METRIC

We are now in a position to introduce our new overall metric. In Table II, each rank (row) has five component scores. Since each pair of components is orthogonal over all persons, the five components can be regarded as providing an orthogonal coordinate system of a five-dimensional Euclidean space, and each rank of persons can be thought of as a point in this five-space.

Let us now inquire about the *distance between each pair of ranks* in this five-space. The Euclidean definition of distance between two points requires us to take the *square root*

of the sum of squares of the differences between the corresponding coordinates of the two points. Thus, in the five-space of Table II, the Euclidean distance between ranks 5 and 4 is $\sqrt{(\cdot 24)}$, since

$$(\cdot 4629-\cdot 2771)^2+(\cdot 2440+\cdot 0488)^2+(\cdot 1179+\cdot 1650)^2+(\cdot 0463+\cdot 1389)^2+(\cdot 0126+\cdot 0630)^2=\cdot 24.$$

For the purposes of our new metric, it turns out to be more profitable to regard the five-space of Table II as a *non-Euclidean* space, by defining the distance between two points to be the *square* of the Euclidean distance: that is, we shall dispense with taking square roots as required by the Euclidean metric. Accordingly, our non-Euclidean distance between ranks 5 and 4 is ·24, in distinction to the Euclidean distance of $\sqrt{(\cdot 24)}$. The non-Euclidean distances between each pair of ranks of Table II are given in Table III.

TABLE III. THE NON-EUCLIDEAN DISTANCES

The Euclidean-Distances-Squared between each Pair of Ranks in the Five-Space defined by the Principal Components in Table II

Rank of Persons	Rank of Persons					
	5	4	3	2	1	0
5	0	·24	·39	·52	·67	·91
4	·24	0	·15	·28	·43	·67
3	·39	·15	0	·13	·28	·52
2	·52	·28	·13	0	·15	·39
1	·67	·43	·28	·15	0	·24
0	·91	·67	·52	·39	·24	0

A striking feature of Table III is immediately apparent. If $i < j < k$, then the distance from rank k to rank i is greater than the distance from rank k to rank j. Thus, the distance between rank 4 and rank 0 is greater than the distance between rank 4 and rank 1 (·67 as compared with ·43). But the distances between adjacent ranks are not necessarily equal; for example, rank 5 is farther from rank 4 than rank 4 is from rank 3 (·24 as compared with ·15).

The properties described in the preceding paragraph would hold also were the Euclidean distance function (or any other monotonely related distance function) to be used. Our particular non-Euclidean distance function has the following distinctive feature: the distance between any two ranks i and k is precisely the sum of the distances from i to j and from j to k, where j is any rank intermediate to i and k. For example, rank 3 is intermediate to ranks 4 and 1; the distance from rank 4 to rank 3 is ·15, while from rank 3 to rank 1 it is ·28. We have ·15 + ·28 = ·43; and ·43 is also the distance between ranks 4 and 1 as computed independently from Table II and recorded in Table III.

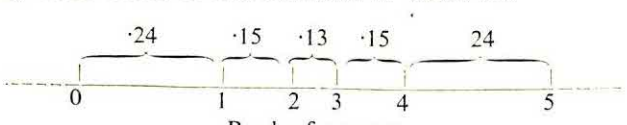


FIG. 1.—The Ranks of Persons Plotted on a Straight Line with Intervals as Defined by the Non-Euclidean Distance Function of Table III.

This additive property of our non-Euclidean distance function implies that we can mark off our ranks of persons as points along a straight line, the distances between the points being proportional to the values given by this distance function, as shown in Fig. 1. All the entries in Table III can be reproduced from Fig. 1. The distance between any two ranks is the sum of the lengths of all intervals between them in Fig. 1.

The new metric portrayed in Fig. 1 maintains the original scale rank order of the types of persons, just as do the previously known three other kinds of metric. It gives distances between the ranks which are different than from those given by the metric that is best in the sense of least squares (the first principal component). The least-squares metric is derived by obtaining projections of the ranks on a best *single* continuum. Our new metric looks at all the ranks as in a complete multi-dimensional space, and does not take projections on only one of the components of that space. We now see that, if we measure distances in a certain non-Euclidean manner, the scale ranks all lie on a ' straight line ' in the resulting non-Euclidean space.

V. POSSIBLE APPLICATIONS FOR NON-SCALABLE UNIVERSES AND FOR QUASI-SCALES

It is not clear at present what further implications the new metric has for the theory of perfect scales as such, or for the psychological interpretations of scales (as from the metric of absolute psychological types discussed in [3]). It may have important applications in the analysis of *non*-scalable data.

The general equations for the principal components of qualitative data hold whether the data form a perfect scale or not, and indeed were derived in [1] before the theory of perfect scales of [2] was developed. *Any* set of qualitative items can be regarded as establishing an n-dimensional coordinate system of principal components for any population of respondents; and each person can be regarded as a point in this space. Tables analogous to Tables II and III above can be computed. Inspecting the distances between the persons can sometimes show that the points lie on some simple surface, if not on a straight line as in the case of a scale. This would give a means of analysing simple non-scale structures. It may turn out that, for some non-scale structures, a different non-Euclidean distance function may be more appropriate than the present function. The new type of metric may also prove helpful in studying quasi-scales. A problem to be explored is the effect on the new metric of errors of reproducibility. It may be that the new metric actually defines a class of quasi-scales, for it may possibly be perfectly additive for data which do not form a perfect scale, according to the mathematical analysis below.

VI. PROOF OF THE ADDITIVE PROPERTY OF THE NEW METRIC

The numerical example worked out above was for a special case of five types of items, and where uniform frequencies (rectangular distribution) were assumed both for the types of persons and the types of items. Different frequencies would yield different principal component scores from those of our example. We shall now show that the additive property holds for any frequency distribution, and for any number of items. While our analysis is the same as for dichotomous items, this entails no real loss of generality; for any item can be reduced to dichotomous form, or expressed as a set of dichotomies.

We shall use the results of [2] and essentially the same notation. As in [2], we may consider a perfect scale of m types of dichotomous items, yielding $m + 1$ types of persons ranked from 0 to m, and which is resolved into a set of $m + 1$ principal components. Let f_i denote the relative frequency of the ith rank of persons ($i = 0, 1, \ldots, m$), and let x_{ai} denote the score of the ith rank on the ath principal component ($a, i = 0, 1, \ldots, m$).

For convenience, the numbering of the principal components will be regarded as being in opposite order of the size of the corresponding latent roots. In particular, $a = 0$ denotes the *constant* principal component (with latent root 1). Since the component scores are determined only up to a constant of proportionality, we can normalise them so that

$$\sum_{i=0}^{m} f_i x^2_{ai} = 1 \qquad (a = 0, 1, \ldots, m). \qquad (1)$$

In [2] it has been proved that all latent roots are distinct, so that any two different principal component score vectors are necessarily orthogonal to each other with weight function f_i. Given also (1), the $m + 1$ components yield an orthonormal set, or

$$\sum_{i=0}^{m} f_i x_{ai} x_{bi} = \delta_{ab} \qquad (a, b = 0, 1, \ldots, m), \tag{2}$$

where δ_{ab} is Kronecker's delta (equals 1 if $a = b$, and 0 if $a \neq b$).

Equations (2) imply that the square matrix of order $m + 1$, $\| \sqrt{(f_i)} x_{ai} \|$, is an orthogonal matrix. This matrix must then be orthonormal by rows as well as by columns, i.e.

$$f_i \sum_{a=0}^{m} x_{ai} x_{aj} = \delta \qquad (i, j = 0, 1, \ldots, m). \tag{3}$$

Let η_a^2 be the correlation ratio squared of the ath principal component. Then $\eta_0^2 = 0$; for the constant component has no variation, although the latent root for this component is 1. Our new metric, as illustrated in sect. IV above, *defines the non-Euclidean distance between two ranks i and j to be* $d(i, j)$, where, recalling we already have normalised the x_{ai} according to (1),

$$d(i, j) = \sum_{a=1}^{m} \eta_a^2 (x_{aj} - x_{ai})^2 \qquad (i, j = 0, 1, \ldots, m). \tag{4}$$

This metric will be additive if, and only if, for any three ranks i, j, and k (where $i \leqq j \leqq k$) the following property holds:

$$d(i, k) = d(i, j) + d(j, k) \qquad (i \leqq j \leqq k). \tag{5}$$

Now, the following identity is obvious:

$$x_{ak} - x_{ai} \equiv (x_{ak} - x_{aj}) + (x_{aj} - x_{ai}). \tag{6}$$

We can square both members, multiply through by η_a^2, sum over a, and use (4) to obtain the further identity

$$d(i, k) = d(i, j) + d(j, k) + 2 \sum_{a=1}^{m} \eta_a^2 (x_{ak} - x_{aj})(x_{aj} - x_{ai}) \qquad (i, j, k = 0, 1, \ldots, m). \tag{7}$$

Clearly (7) will reduce to (5) if, and only if, the last member of (7) vanishes whenever $i \leqq j \leqq k$, or

$$\sum_{a=1}^{m} \eta_a^2 (x_{ak} - x_{aj})(x_{aj} - x_{ai}) = 0 \qquad (i \leqq j \leqq k). \tag{8}$$

Thus our task will be ended if we prove that (8) always holds for a perfect scale.

As a preliminary to the proof, we establish a sufficient condition for (8). Let Δx_{ai} be defined as the difference

$$\Delta x_{ai} = x_{a, i+1} - x_{ai} \qquad (i = 0, 1, \ldots, m - 1). \tag{9}$$

Differencing will always be with respect to the second subscript, so instead of $\underset{i}{\Delta}$, we shall write Δ throughout. If $j > i$, then we can write, using (9),

$$x_{aj} - x_{ai} = \sum_{s=0}^{j-i-1} \Delta x_{a, i+s} \qquad (j > i). \tag{10}$$

Similarly, if $k > j$, we can write

$$x_{ak} - x_{aj} = \sum_{t=0}^{k-j-1} \Delta x_{a, j+t} \qquad (k > j). \tag{11}$$

Using (11) and (10) in (8), we can write

$$\sum_{a=1}^{m} \eta_a^2 (x_{ak} - x_{aj})(x_{aj} - x_{ai}) = \sum_{s=0}^{j-i-1} \sum_{t=0}^{k-j-1} \left(\sum_{a=1}^{m} \eta_a^2 \Delta x_{a,\, i+s} \Delta x_{a,\, j+t} \right) \qquad (i < j < k). \qquad (12)$$

Let p_{ij} be defined by:

$$p_{ij} = \sum_{a=1}^{m} \eta_a^2 \Delta x_{ai} \Delta x_{aj} \qquad (i, j = 0, 1, \ldots, m-1). \qquad (13)$$

Clearly, a *sufficient* condition for the righthand member of (12) to vanish is that $p_{i+s,\, j+t} = 0$ for all s and t in the given ranges for the respective i, j, and k. Surely, then, a sufficient condition for the right member of (12) always to vanish is that:

$$p_{ij} = 0 \qquad (i \neq j;\ i, j = 0, 1, \ldots, m-1). \qquad (14)$$

If we prove (14), we shall have proved (8) via (12), and hence (5) via (7). The left member of (8) is identically zero if $i = j$ or $j = k$, and hence we need consider the case only where $i < j < k$, as we shall now do via (14).

For the proof of (14), we shall use the *difference equations* of the perfect scale derived in [2]. These equations can be written in the following form, suitable to our present purposes [cf. 2, p. 345, equations (51)-(53)]:

$$c_0 \Delta x_{a0} = -\frac{1}{\eta_a^2} - f_0 x_{0a} \qquad (a = 1, 2, \ldots, m) \qquad (15)$$

$$c_{i+1} \Delta x_{a,\, i+1} - c_i \Delta x_{ai} = -\frac{1}{\eta_a^2} f_{i+1} x_{ai,\, i+1} \qquad \begin{pmatrix} i = 0, 1, \ldots, m-2 \\ a = 1, 2, \ldots, m \end{pmatrix} \qquad (16)$$

$$c_{m-1} \Delta x_{a,\, m\ 1} = \frac{1}{\eta_a^2} f_m x_{am} \qquad (a = 1, 2, \ldots, m). \qquad (17)$$

We have modified the notation slightly from that in [2]. Our present c_i are n/N times those of [2, equation (34)], and our $1/\eta_a^2$ are n/N times the ϕ of [2, equation (35)], except for when $a = 0$. It is essential to note that $1/\eta_0^2$ is *not defined* in our present notation, because $\eta_0^2 = 0$, although ϕ_0 *is* defined as $\phi_0 = 0$ in [2, pp. 345 f.]. Our present equations (15) through (17) are not stated for the constant component, to which we have to give special attention in the algebra that follows.

Let x_0 be the common constant value of the x_{0i} ($i = 0, 1, \ldots, m$). Under condition (1) above, x_0 is fixed by $x_0^2 \sum_{i=0}^{m} f_i = 1$. Explicitly, if $\sum_{i=0}^{m} f_i = N$ as in the notation of [2, equation (4)], then $x_0 = \sqrt{(1/N)}$. From (3), if we take the case $a = 0$ out of the summation and write it separately, we can state for $i \neq j$ that

$$f_i \sum_{a=1}^{m} x_{ai} x_{aj} = -f_0 x_0^2 \qquad (i \neq j;\ i, j = 0, 1, \ldots, m). \qquad (18)$$

Taking the first difference of both members of (18) with respect to j, and dividing by f_i, shows that

$$\sum_{a=1}^{m} x_{ai} \Delta x_{aj} = 0 \qquad \left(i \neq j, j+1; \quad \begin{matrix} i = 0, 1, \ldots, m \\ j = 0, 1, \ldots, m-1 \end{matrix} \right). \qquad (19)$$

Now, if we multiply both members of (15) through by $\eta_a^2 \Delta x_{aj}$, sum over a, and use notation (13), and equation (19), we see that, since no c_i can vanish,

$$p_{ij} = 0 \qquad (j = 1, 2, \ldots, m-1). \qquad (20)$$

Multiplying both members of (16) through by $\eta_a^2 \Delta x_{aj}$ and summing over a will similarly show that

$$c_{i+1} p_{i+1,\, j} - c_i p_{ij} = 0 \qquad \left(i \neq j, j+1; \quad \begin{matrix} i = 0, 1, \ldots, m-2 \\ j = 0, 1, \ldots, m-1 \end{matrix} \right). \qquad (21)$$

Stating (21) for $i = 0$, and using (20), now shows that

$$p_{ij} = 0 \qquad (j = 2, 3, \ldots, m-1). \tag{22}$$

Stating (21) for $i = 1$, and using (22), etc., shows that in general

$$p_{ij} = 0 \qquad \left(j > i; \quad \begin{matrix} i = 0, 1, \ldots, m-2 \\ j = 0, 1, \ldots, m-1 \end{matrix}\right). \tag{23}$$

But $p_{ij} \equiv p_{ji}$ from (13). Hence (23) is equivalent to (14), and we have completed the proof of the additive property (5) of our new metric.

It may be remarked that property (14) shows that, from (13), $\| \eta_a \Delta x_{ai} \|$ is a square matrix of order m that is orthogonal by columns. The Δx_{ai} are closely related to the principal components *of the categories* of the items in the scale, as shown by [2, equation (30)]; the weight function required for orthogonalization by rows is correspondingly implied by [2, equation (94)].

A full treatment of the related problem of category components, when items are not dichotomous but quantitative variables, is given in [4].

REFERENCES

1. Guttman, Louis. 'The quantification of a class of attributes.' In Horst, *et al.*, *The Prediction of Personal Adjustment*. New York: Social Science Research Council, 1941; pp. 319-348.
2. Guttman, Louis. 'The principal components of scale analysis.' In Stouffer, *et al.*, *Measurement and Prediction*. Vol. IV of *Studies in Social Psychology in World War II*. Princeton University Press, 1950 ; Chapter 9, pp. 312-361.
3. Guttman, Louis. 'The principal components of scalable attitudes.' In Lazarsfeld, *et al.*, *Mathematical Thinking in the Social Sciences*. The Free Press, 1954.
4. Guttman, Louis. 'A generalized simplex for factor analysis.' *Psychometrika*, XX, 1955 (in the press).

Vol. VIII Part I The British Journal of Statistical Psychology May 1955

THE CORRELATION BETWEEN VARIATE-VALUES AND RANKS IN SAMPLES FROM DISTRIBUTIONS HAVING NO VARIANCE

By ALAN STUART
Division of Research Techniques, London School of Economics

1. In [2], it was shown that in samples of size n from a continuous distribution function $F(x)$ with mean μ and variance σ^2, the correlation between variate-values and the corresponding ranks is given by

$$C_n = \left\{\frac{12(n-1)}{(n+1)}\right\}^{\frac{1}{2}} \cdot \frac{\left\{\int xF(x)\,dF(x) - \tfrac{1}{2}\mu\right\}}{\sigma}. \tag{1}$$

We now consider the problem in the case when the distribution has no variance.

Moments can fail to exist only if the range of the distribution is either singly- or doubly-infinite. In the first place, we consider only a portion of the distribution defined on an arbitrarily wide range (a, b). For this interval, (1) holds, with a and b as the limits of integration in $\int xF\,dF$ and in the integrals defining μ and σ^2.

2. Consider first the case of a singly-infinite range, say upwards. We may take a to be the lower terminal of the distribution, and take the limit of (1) as $b \to \infty$. If we write (1) as

$$C_n = K \,.\, R, \tag{2}$$

we have

$$\operatorname{Lim} R^2 = \operatorname{Lim} \frac{\left\{\int_a^b xF\,dF - \tfrac{1}{2}\int_a^b x\,dF\right\}^2}{\left\{\int_a^b x^2\,dF - \left(\int_a^b x\,dF\right)^2\right\}}. \tag{3}$$

If the mean of the distribution exists, the numerator of (3) has a finite limit; for $\int xF\,df$ converges whenever $\int x\,dF$ does, since $0 \leqslant F \leqslant 1$. And since the denominator of (3) diverges by hypothesis, the ratio R^2 tends to zero.

The problem becomes non-trivial when the mean, as well as the variance, does not exist. We then apply L'Hôpital's rule to (3), differentiating numerator and denominator to obtain

$$\operatorname{Lim} R^2 = \operatorname{Lim} \frac{2(F(b) - \tfrac{1}{2})\left\{\int_a^b xF\,dF - \tfrac{1}{2}\int_a^b x\,dF\right\}}{\left\{b - 2\int_a^b x\,dF\right\}},$$

and on repeating the operation,

$$\operatorname{Lim} R^2 = \operatorname{Lim} \frac{2f(b)\left[b(F(b) - \tfrac{1}{2})^2 + \left\{\int_a^b xF\,dF - \tfrac{1}{2}\int_a^b x\,dF\right\}\right]}{1 - 2bf(b)}. \tag{4}$$

Now if there exists a value of x, say X, such that

$$f'(x) < 0 \quad \text{for all} \quad x > X,$$

i.e., such that the frequency function is monotone, decreasing beyond some point, it will follow from a theorem quoted in [1], p. 124 that, even though the mean does not exist,

$$\text{Lim } bf(b) = 0. \tag{5}$$

Substituting (5) in (4), we find

$$\text{Lim } R^2 = \text{Lim } \frac{2\left\{\int_a^b xF\,dF - \frac{1}{2}\int_a^b x\,dF\right\}}{1/f(b)}. \tag{6}$$

Applying L'Hôpital's rule again to (6), we obtain

$$\text{Lim } R^2 = \text{Lim } \frac{-2\{F(b) - \frac{1}{2}\}b\{f(b)\}^3}{f'(b)}, \tag{7}$$

and since

$$\text{Lim } F(b) = 1,$$

(7) becomes

$$\text{Lim } R^2 = -\text{Lim } \frac{b\{f(b)\}^3}{f'(b)},$$

$$= -\text{Lim } bf(b) \cdot \frac{\{f(b)\}^2}{f'(b)}. \tag{8}$$

From (5), using L'Hôpital's rule again,

$$0 = \text{Lim } \frac{b}{1/f(b)} = -\text{Lim } \frac{\{f(b)\}^2}{f'(b)}. \tag{9}$$

Putting (5) and (9) into (8), we have, using (2),

$$\text{Lim } C_n = \text{Lim } R^2 = 0. \tag{10}$$

An example of the type of distribution discussed is

$$dF(x) \propto \frac{dx}{x^2}, \qquad (0 < x \leqslant \infty)$$

which has no mean since $\int_0^\infty \frac{dx}{x}$ diverges.

3. The same result holds for a doubly-infinite range if we write $a = -b$ and interpret the integrals as principal values. For since

$$\frac{d}{db}\int_{-b}^{b} g(x)\,dx = 2\frac{d}{db}\int_c^b g(x)\,dx,$$

the effect of having a doubly infinite range is simply to introduce a factor 2 whenever an integral is differentiated, and this does not change the value of the limit at (10). Thus, for example, the Cauchy distribution

$$d \quad F(x) \propto \frac{dx}{1 + x^2}, \qquad (-\infty \leqslant x \leqslant \infty),$$

which has no moments and its monotone in is tails, has $C_n = 0$ by (10). We conclude that, for any continuous distribution with no moments which is ultimately monotone,

the correlation between variate-values and ranks is zero. However, this property does not characterize such distributions, for in [2] it was shown that for Gamma variates with parameter m (>0),

$$\lim_{m \to 0} C_n = 0,$$

and that generally, when the variance exists, a necessary condition for $C_n = 0$ is that $f(x)$ be unbounded at some point within its range.

REFERENCES

1. Knopp, Konrad (1928). *Theory and Application of Infinite Series.* London and Glasgow: Blackie.
2. Stuart, Alan (1954). 'The Correlation between Variate-values and Ranks in Samples from a Continuous Distribution.' *Brit. J. Statist. Psychol.*, VII, 37-44.

ERRATA

The following corrections should be made in the article on 'The Correlation between Variate-values and Ranks in Samples from a Continuous Distribution,' this *Journal*, VII, Pt. I, pp. 37-44.

1. P. 40, eqn. (22): Right side should be $1/(2.3^{\frac{1}{2}})$, not $1/(2.3)^{\frac{1}{2}}$.
2. P. 41, below eqn. (24): delete 'a result . . . much earlier.'
3. P. 42, end of Section V should continue: 'The result (24) for r in the normal case has been given by Burt (1953), who obtained it earlier.'
4. P. 43, Example 1, Line 5: delete 'in virtue of section V.'
5. P. 43, Example 1, Lines 6 and 7, delete: 'asymptotically' (twice).
6. P. 43, Example 1, second paragraph, Line 2: The result for asymptotic relative efficiency should read $(C^2)^{\frac{1}{2}} = (3/\pi)^{\frac{1}{2}}$, not $C^2 = (3/\pi)$ as printed.

ALAN STUART.

Vol. VIII Part I The British Journal of Statistical Psychology May 1955

A NOTE ON THE USE OF BURT'S FORMULA FOR ESTIMATING FACTOR SIGNIFICANCE

By EDWARD E. CURETON

Burt [1] gives as an approximate formula for the standard error of a centroid or simple-summation factor loading :

$$\sigma_a = (1-a^2)\sqrt{\{n/[N(n-s+1)]\}}, \quad . \quad . \quad . \quad . \quad . \quad (1)$$

where a is the loading, N is the number of subjects in the sample, n is the number of tests or variables, and s is the ordinal number of the factor.

Vernon suggests that the sth factor may be considered insignificant whenever at least half its loadings fall short numerically of twice their standard errors.[2] It is not necessary to compute the standard errors of all n factor loadings to apply this test. If we examine (1), we see that the expression under the radical is the same for all loadings on a given factor.

Let $a/\sigma_a = k$ (according to Vernon's recommendation, k should be 2); then $a = k\sigma_a$, and from (1),

$$a = (1-a^2)\,C, \quad . \quad . \quad . \quad . \quad . \quad . \quad (2)$$

$$\text{where } C = k\sqrt{\{n/[N(n-s+1)]\}}. \quad . \quad . \quad . \quad . \quad . \quad (3)$$

Now (2) is a quadratic in a, whose positive root is

$$a' = [-1+\sqrt{(1+4C^2)}]/2C. \quad . \quad . \quad . \quad . \quad . \quad (4)$$

If we find C from (3) and then a' from (4), a' is the critical value of a: all loadings numerically greater than a' are significant and all loadings numerically less than a' are insignificant at the $k\sigma$ level.

EXAMPLE:

$$N = 600,\ n = 30,\ s = 11,\ k = 2.$$

$$C = 2\sqrt{\{30/[600\,(30-11+1)]\}} = \cdot 1.$$

$$a' = [-1+\sqrt{(1\cdot 04)}]/\cdot 2 = \cdot 099.$$

The eleventh factor will be significant, according to Vernon's standard, if and only if more than half its loadings are numerically greater than ·099.

[1] C. Burt and C. Banks. *Annals of Eugenics*, XIII, 255; also this *Journal*, 'Tests of Significance in Factor Analysis', v, 121, eqn. xxvi.

[2] Burt's criterion is more lenient : with borderline cases at least one quarter must exceed twice the standard error for the factor to be significant. But in general he also takes the size of the loading into account.

Vol. VIII Part I The British Journal of Statistical Psychology May 1955

A PSYCHOLOGICAL STUDY OF TYPOGRAPHY

By CYRIL BURT, W. F. COOPER, and J. L. MARTIN

University College, London

Abstract. A wide variety of type-faces have now become available for books and journals. It therefore seemed desirable to investigate (*A*) the legibility and (*B*) the aesthetic merits of those in more frequent use.

(*A*) Using tests of speed and comprehension, we have studied the influence of type-face, boldness, size, interlinear spacing, length of line, and width of margin on legibility both with children and with adults. The results have furnished provisional norms for children's reading books and for scientific journals such as the present.

(*B*) Factorial methods, supplemented by an analysis of introspections, appear to yield a classification of both readers and type faces based on aesthetic preference; and the data incidentally obtained throw considerable light on the reasons for such preferences.

I. PREVIOUS INVESTIGATIONS OF TYPOGRAPHICAL PROBLEMS

Printing as a Medium of Communication. Psychologists have of late shown increasing interest in what is called 'the theory of telecommunication', that is, communication between persons who are not within visible or auditory range of each other.[1] The statistical treatment of the issues involved, which is the most distinctive feature of the 'theory', has been developed primarily in connection with telephonic engineering; and, perhaps largely for this reason, the attention of the psychologist has been concentrated mainly on auditory methods of transmission. But communication by visible symbols is not only the oldest but by far the commonest method of imparting information to a distant recipient; and the problems of setting up print and reading print—the most thoroughly standardized of all such devices—are equally amenable to the same mode of analysis.

If we take the 'phoneme'[2] as the unit of information, the English language presents us with a fairly well defined ensemble of alternative elements—12 vowels and 37 consonants—which the printer proceeds to encode by using an alphabet of 26 visible signals, each of which may have four different forms, roman, italic, upper case (or capital), and lower case. Analogous to the audible 'noise' which interferes with the accurate transmission of the audible information, there is a visuo-mental interference or 'blur', which impedes the accurate transmission of visible information. Accordingly, just as the intelligibility of a wireless message is measured by a logarithmic function of the signal-to-

[1] Cf. Burt, this *Journal*, IV, pp. 195-200: also D. MacKay, 'Quantal Aspects of Scientific Information', *Phil. Mag.*, LXIII (1950), G. A. Miller, *Language and Communication* (New York, McGraw-Hill, 1951), and C. E. Shannon, 'Communication Theory: Exposition of Fundamentals' (London, Ministry of Supply, 1953).

[2] Otto Jespersen, *Lehrbuch der Phonetik* (1916), Daniel Jones, *Outline of English Phonetics* (1928). Any phonetic symbol, ε (for example), represents, not an invariable speech-sound made with the vocal organs always in the same place, but a family of ε-sounds varying slightly (even with the same speaker) according to the vocal context. Such a family is termed a 'phoneme'. The neglect of this fact is a common source of trouble in teaching children, particularly backward children, to read: see Burt, *The Backward Child*, Appendix IV and refs. It may be noted that three more phonemes are needed to represent standard American speech.

noise ratio, so the intelligibility of a printed message could be measured by a similar function of what we might call the symbol-to-blur ratio.[1]

As was perhaps only natural for the engineer, the formulae which he has developed to measure information treat the efficacy of the signalled message as a kind of 'energy' or capacity for doing work, and its negative is represented by an expression similar to that adopted by Boltzmann in formulating 'entropy' in terms of probability. However, for our present purpose it seems desirable to take account of yet another variable characteristic. Of two media enabling the same message to be communicated to the same recipient,with the same accuracy, that which delivers the information the more speedily must plainly be regarded as the more effective. The consideration of speed of reading therefore adds a further complication to the problem, and should perhaps lead us to look for some formula which will enable us to measure the 'power' of the media.

We believe that the adoption of the concepts and methods employed in the theory of communication would greatly increase the logical rigour of the psychologist's investigations. But we are somewhat doubtful whether, in the existing stage of knowledge, the use of the corresponding algebraic formulae will be of much service. Psychological measurements are so crude and inexact that the elaborate mathematical refinements of these newer techniques seem out of place. In any case, the following article is intended merely as a preliminary exposition; and it has been our aim to make it comprehensible to teachers and others with no special knowledge of advanced statistical theory. For the most part, therefore, we shall for the present keep to the simplest possible methods of assessment; and measurements of accuracy and speed will be reported separately in terms of units of time and of ideational items. We shall begin by investigating the ways in which the visible appearance of printed matter aids or hinders the communication of the ideas which it seeks to express, in a word, the conditions affecting *legibility*.

II. LEGIBILITY

The Legibility of Children's Reading Books. In this country the typographical questions which have chiefly attracted the attention of the psychologist relate to the sizes and shapes of the letters used for printing children's reading books. Even on these points the number of British investigations has been surprisingly few; and those who are concerned with the selection of suitable type for reading books or for verbal tests are still guided (so far as they pay any consideration to the matter at all) by recommendations put forward nearly forty years ago.

In 1912 a committee was appointed by the Education Section of the British Association to investigate 'the influence of school books upon eyesight'. In order to secure evidence on certain of the questions then raised, Burt (at that time Psychologist in the Education Department of the London County Council) and James Kerr (then Medical Officer in the same Department) carried out a number of investigations among children in London schools. It was agreed that Dr. Kerr should deal with the clinical aspects, while Burt undertook a series of experiments in the classroom and in the laboratory. The Council's printer generously supplied us with a wide variety of test-sheets printed to our own specifications.

The results of our investigations were summarised in the L.C.C. Annual Reports and later in Dr. Kerr's larger treatise [7].[2] The Final Report of the Committee of the British Association [2] was published in 1917; and included a tabulated set of standards showing the size of type and the style of printing suited to pupils of different ages. Most of the school books since published by the more competent firms have conformed fairly closely with the standards then laid down.[3] During

[1] One of our chief methodological difficulties has been to decide whether to take the letter or the word (and their respective significations) as our unit. In applying the theory of communication to the study of the child's speaking vocabulary, Burt preferred to take the word as unit (cf. this *Journal*, *loc. cit. sup.*, p. 199 and refs.); and, as will be seen later, we have in the present series of experiments kept mainly to words or ideational units. Most psychologists who have investigated legibility have limited themselves to the legibility of isolated letters. But with modern methods of teaching children to read, the legibility of the word as a whole is even more important than the legibility of the constituent letters. The two modes of measurement do not always yield the same result. A complete investigation requires the use of both.

[2] Cf. also Burt [3], pp. 271, 340-1. Dr. Kerr's standards for size and interlinear space are not quite so generous as those of the Brit. Assoc. Committee. A review of practices generally current before these inquiries will be found in a paper by Mr. G. F. Daniell (Chief Examiner to the London County Council) in the *Annual Report* of the Brit. Assoc. for 1911, p. 633. After the first World War, a suggestive monograph was published by the Stationery Office (L. A. Legros, *Note on the Legibility of Printed Matter*, 1922; cf. also [4]). See also R. D. Morss, 'The Neglected School Book', *Monotype Recorder*, XXXIV, 1935, pp. 5f.

[3] A fairly comprehensive summary of the chief American investigations will be found in [22], chap. IV (pp. 101-25; see also bibliography, pp. 452-74). But the reader who has no easy access to the original articles will at times be disappointed to find that the age of the subjects of the experiments is seldom stated. Occasionally, too, the data are said to yield significant differences between two or more different modes of printing, but the authors omit to staet in what direction the differences pointed.

the next ten years a series of laboratory researches, often involving highly skilled techniques, were carried out in America [9-14]. But since the Second World War little fresh work has apparently been attempted in either country. Owing to the great advances in printing during the last two or three decades, and the wide variety of types meanwhile introduced by the Monotype Corporation and other foundries, it seems highly desirable that the problems should be taken up afresh.

A few years ago, when the British Psychological Society proposed to establish a journal of statistical psychology, Burt, with the help of several research students, began a series of preliminary investigations on legibility and typographical preferences among educated adults ; and the more recent decision of the Society to reconsider its choice of printer has furnished an additional reason for ascertaining the views of potential readers and subscribers to this journal. The appeal inserted in the last issue has elicited numerous comments and replies ; and we should like to thank our correspondents for their help. With the aid of teachers and others, we have continued and considerably extended the earlier experiments on school children. The account that follows will summarize, as succinctly as possible, the main conclusions that have so far emerged. As so often happens, one of the most valuable results has been to show where the methods adopted were at fault and to indicate more effective lines of inquiry.

Methods. To determine the legibility of type, a large variety of methods have been employed by previous investigators. The commonest may be broadly classified as follows :

1. Ease of reading letters, words, or sentences, judged by the distance at which they can be read.
2. Accuracy of reading letters or words with brief tachistoscopic exposures.
3. Speed of reading passages of prose, when the reader's aim is to grasp the content of the passages.
4. The observation of eye-movements, eye-blinking, and other objective symptoms.

There can, we think, be little doubt that the results obtained often depend very largely on the experimental procedure used. For example, types that appear more legible than others when tested by the distance method may prove less legible when tested by speed of reading.

We ourselves found the first two methods quite effective for investigating the legibility of isolated letters or figures, particularly where the shape is apt to differ according to the style of type. For our main purpose, however, which was to determine the legibility, not of isolated words or characters, but of consecutive sentences, such procedures proved inappropriate. With children the study of eye-movements, fixation-time, blinking, and the like was at times informative, especially in the case of young or backward readers, and of those who suffered from minor visual defects. But on the whole, its value was merely supplementary ; it might often explain why certain styles of printing were unsuitable, but it did not afford a practicable basis for systematic comparison. The method therefore on which we have principally relied consisted in getting each reader to read silently a page of simple consecutive prose, timing his speed, and then determining by means of a questionnaire how much of the details had been grasped.

For the preliminary experiments children were tested in class, and a fixed time-limit was imposed. For the final experiments individual testing was adopted, and proved far superior to group testing. Every child was asked to read aloud an entire page, which had been devised to form a unit complete in itself ; the time required was measured with a stop-watch, and the questions answered orally.[1] At the same time a careful look-out was kept for pauses, irregular eye-movements, and other signs of difficulty or fatigue.

In the later stages of our work we introduced an important modification. We became gradually convinced that short tests, such as have been used by many previous investigators, can often be quite uninformative and at times positively misleading. In changing to a fresh or unfamiliar type-face, the reader needs time to become adequately adapted. Moreover, differences in legibility and fatiguability may not appear during the first few minutes : many readers who can read 8-point as accurately and as quickly as 10-point, provided only a brief paragraph has to be read, often show signs of strain with the smaller types after they have been reading one or two pages. Accordingly, instead of handing the reader the essential test-passage straightaway, we first gave him a preliminary booklet printed in the same style of type and kept him reading it for 5 minutes, before subjecting him to the final test.

[1] In the earlier experiments much of the testing was carried out by Mr. T. R. Kingsley (who had had considerable experience in an Institute where printing and book-making were regular subjects) and by Miss Violet Pelling. In the more recent experiments most of the testing has been carried out by Mr. W. F. Cooper and by Miss J. L. Martin, who therefore appear as joint authors.

Influence of (a) *Type Faces.* The legibility of a page of print is the resultant of many different factors—the size, the form, the thickness or ' boldness ' of the letters, the width of the line, the distance between successive lines, the texture of the paper, and the intrinsic interest of the subject-matter itself. We have carried out experiments on each of these variables in turn.

We began by trying to assess the relative legibility of the different book-faces that are now in common use. For this the experimental plan ultimately adopted was similar to that described in the investigation on reading carried out in conjunction with Mr. B. Lewis.[1] The method then used was virtually a combination of randomization with a scheme of Latin squares ; but certain minor modifications were introduced to meet the special requirements of psychological research. For the present inquiry ten groups, each consisting of 15 boys and 15 girls aged approximately 10·0 to 11·0 years, were selected by means of preliminary tests, the members being carefully matched so that the means and standard deviations of each group were approximately the same as regards both general intelligence and reading ability. Thirty short stories, each about a page in length, were constructed, and set up in 10 different type-faces—all roman—with 3 sizes for each (9-, 10-, and 11-point for the larger faces, 10-, 11-, and 12-point for the smaller). For each passage a questionnaire of 25 questions was drawn up ; and the drafts for both stories and questions were tried out first of all, and where necessary amended, until they appeared to be about equal in difficulty. In the final experiment every pupil was asked to read 10 passages, one in each style of type, the different stories and the different sizes being allotted according to the randomized scheme.[2] The preparatory trials had shown that, after the first test had been taken, the order in which such passages were read made little difference to the accuracy of the reading. Accordingly, the main experiment started with a preliminary story, used merely to provide a certain amount of practice. A short series of supplementary experiments were made with italic type-faces ; but for these the groups were somewhat smaller.

TABLE I. SPEED AND COMPREHENSION WITH DIFFERENT TYPE FACES

Type Face	Roman		Italic	
	Time (*Seconds*)	Comprehension (*Items*)	Time (*Seconds*)	Comprehension (*Items*)
Old Style (161)	96	28·0	—	—
Imprint (101)	97	29·1	113	26·9
Times New Roman (327)	102	27·9	118	25·8
Plantin (110)	105	26·8	129	25·0
Bembo (270)	106	26·5	135	23.7
Baskerville (169)	108	27·0	137	24·8
Caslon (128)	112	27·6	138	25·0
Scotch (46)	114	25·9	108	27·3
Modern Extended No. 1 (7)	116	25·8	116	26·1
Bodoni (135)	125	25·1	124	25·8

The numbers in brackets refer to the serial numbers of the Monotype Corporation. The chief differences between the type-faces used are illustrated by the specimen letters printed in the appendix (pp. 45-57).

Most investigators (cf. 22, pp. 114 and refs.) describe the size of the type used for their tests in terms of ' points ' ($\frac{1}{72}$ in.). But the x-height of different faces of the same nominal size in points differs considerably : for example, with 10-point that of Times New Roman is 0·062 in. ; that of Bembo only 0·050 in. To allow for these divergences the figures obtained with different point-sizes were weighted appropriately and then averaged, so as to reduce all to terms of the same x-height, namely, 0·060 in. (that of 11-point Imprint and Modern 7). The averages so obtained, for both speed and accuracy, are summarized in Table I.

[1] *Brit. J. Educ. Psychol.*, XVI, 1946, pp. 116-32. The reasons for the modifications are there explained in the footnotes on pp. 121 f. ; cf. also *ibid.*, IX, 1939, p. 259.

[2] Most of our reading-passages have been set in monotype. But in our earliest experiments a good deal was either hand set or composed in linotype. In the earlier stages we also used Modern No. 1, Series 1, and Old Style, Series 2—two of the oldest monotype faces, cut in 1900 and 1901 respectively. For these we afterwards substituted Modern Extended No. 1, Series 7, Imprint, and Old Style, Series 161. Times New Roman did not become available until 1933, when *The Times* released the design for general use.

Experiments were also made with Garamond, Fournier, Ehrhardt, and Gill Sans : but for children these were found to be much inferior to the eight type-faces eventually selected.

The data was subjected to an analysis of variance in the usual way. The variance of the means of the different groups, and the variances of the means for the different passages, were statistically non-significant; the variance of the means for the different type-faces was fully significant. The differences between adjacent pairs are too small to be significant statistically; but it will be observed that the order of the type faces in all four columns shows a fair amount of agreement.

The British Association Committee recommended old faces rather than modern, and gave Caslon the preference. In our own experiments we found both Old Style and Imprint far more satisfactory than Caslon. (Old Style and Imprint may be regarded as modernized versions of Caslon, with its more peculiar features removed—notably the sprawling capitals which tend to distort the familiar word-form.) As regards roman types (i.e., the upright letters now ordinarily used for consecutive prose), Imprint appears on the whole to be the best; Plantin [1] and Times New Roman are nearly as good, although with the younger children their short ascenders and descenders seem at times to render the word-form less distinctive.[2]

The modern type faces were the least legible.[3] As the reader will see if he compares 'n' and 'u' in (say) Garamond and Bodoni respectively, the older faces accentuate those parts that are different, while the modern faces accentuate those that are similar. With the latter, therefore, letters are far more readily confused. Legibility is still further diminished by the excessive disparity between the thick vertical strokes and the thin curves and slanting strokes: to a child who is slightly hypermetropic or astigmatic, a word like 'minimum' printed in Bodoni looks like a succession of i's, and the whole line gives the impression of a palisade. Condensed faces (such as are seen in many French reading books) are also liable to look blurred to the hypermetropic eye; on the other hand, excessively expanded faces (such as are occasionally found in American books) tend to disrupt the word form.

Some of the commoner type-faces present special difficulties to children because of the unusual shapes of certain letters. The lower serif or beak on the capital C in Bodoni, Baskerville, and italic Caslon, causes many to mistake it for G. Capital Q in Imprint and Times Roman is sometimes mistaken for O. The old-fashioned Y in italic Caslon and Baskerville puzzled many young readers. The barred italic *J* of Caslon, Imprint, Plantin and Baskerville was read by many children as 'f'; and the italic *h* of Garamond and one or two other old faces tends to be read as '*b*'. In Caslon italic the extreme slope of the letters and the wide separation of many capital initials from the rest of the word also caused confusion. On the other hand, we found that, particularly with very young children, the modern face R (with its curling leg) and Q (with its tail beginning inside the counter) were more readily recognized than the corresponding old face capitals. The larger eye of the modern e makes for greater legibility: children with imperfect vision sometimes mistake a Baskerville or Caslon e for c. The modern faces have one further advantage frequently noted by teachers: the ranging numerals are said to be much easier for children (those of Scotch Roman and Times Roman appear to be easiest of all); whereas the ascending and descending numerals of the older faces occasionally produce an appreciable hesitation, particularly when the figures are grouped.[4] Here as elsewhere, however, much depends upon the style and shapes to which the child has been accustomed. For young children, we are firmly convinced, there would be a great advantage in selecting a single legible set of letter-shapes, keeping strictly to these, and allowing no picturesque divergences. For them too, displayed material, such as titles of stories, headings to chapters, and the like, should be set, not in capitals, but in lower case with capital initials.

We attempted similar experiments with adults (chiefly students and other educated readers) confining ourselves for the most part to seven type-faces only (Imprint, Times, Baskerville,

[1] A smooth hard paper was used: otherwise, owing to its heavier impression, Plantin would have proved the most legible of all: (see section on 'Boldness').

[2] A fount of Times Roman with long descenders is now available. We may add that, although the long ascenders and descenders adopted by most 'modern' faces tend to increase the legibility of the isolated letters for younger children, they spoil the characteristic unity of the words for older readers.

[3] Of the semi-modern type-faces, 'Century' (a type used for a popular series of school readers) appears on the whole to be the most readable. But its merits were not great enough for us to include it in our final selecting a single legible set of letter-shapes, keeping strictly to these, and allowing no picturesque series (see *Monotype Recorder*, XXXIV, 1935, pp. 18 f., and David Thomas, 'School Books and their Typography,' *Printing Review*, XIII, 1934, pp. 5-8).

[4] Several of the 'old faces', e.g., Imprint and Bembo, are now provided with an alternative set of ranging figures. We believe, however, that the question has generally been oversimplified and the effect of habit usually overlooked. Certainly the descending 3 and 5 of Plantin and Imprint are easily confused. On the other hand, with lining figures 3 and 8 are confused, especially in reading small-type mathematical tables. Thus, we are inclined to think that, for tables like those of Chambers, a slightly ascending 6 and 8 would aid discrimination, especially when there is a chance of the plates getting worn. In our opinion, however, the whole subject calls for further investigation by up-to-date psychological techniques.

Bembo, Old Face, Scotch, and Modern). Unfortunately, it was impossible to plan the research as systematically as before or to test such large numbers. The results are therefore not worth reporting in any detail. With adults, we found 10-point Times Roman proved far more legible than 10-point Bembo or Caslon ; but this is mainly due to the fact that with the former the x-height is much larger. When allowance is made for differences in visible size, there is, for persons of normal eyesight, virtually no difference in legibility between the commoner book faces. For older persons and for those with visual defects, type faces varying widely in the thickness of their lines and (to a less extent) those with hardly any variation at all seem the least legible. For such persons the general order of legibility appears to be much the same as for children. The chief exception is Modern No. 1 which causes far less difficulty to adults, particularly if (like Science students) they are already fairly accustomed to it.

(b) *Boldness.* For readers who are hypermetropic or astigmatic, legibility is improved by increasing the heaviness [1] of the type. But the optimal thickening is severely limited—a point which has too often been overlooked : excessive thickening tends to reduce the size of the ' counters ' (i.e. the white inner spaces) in such letters as ' a ' and ' e '. A moderately bold or heavy type seems definitely preferable in books intended either for the very young or the very elderly. For the normal eye of youth there is little gain but rather the reverse.

In several type faces the boldness is achieved mainly by thickening the vertical strokes, as for instance in Times Bold (328) : Times New Roman Semi-Bold (421) avoids these disproportionate differences, but the set is a little too narrow especially in the larger sizes. With children Plantin (110) owes its legibility partly to the fact that it leaves a somewhat darker impression, especially on uncoated paper : but the peculiarities of the capital C and W, and the splayed capital M, as well as the lower case j, are unsuited to the younger readers. For them the most legible fount would seem to be a semi-bold ' Old Style,' e.g., Monotype Old Style Antique (161)—a type design [2] which now seems to be seldom used for ordinary book work. For elderly readers we found Veronese (59)—one of the earliest of the machine type-faces to aim at aesthetic quality [3]—apparently the least tiring.

(c) *Size.* Our next problem was to investigate the influence on legibility of size. We began with Old Style Antique. However, in producing books for younger children printers and publishers still use a wide variety of type faces ; and we therefore decided that norms stated in terms of one of the more widely available founts, with no marked leanings toward any exceptional characteristics either in design or in heaviness, would be more useful than norms stated in terms of a type that is now comparatively infrequent. Accordingly, after some consultation with teachers and publishers, we eventually determined to base our norms on Times New Roman.

The British Association Committee gives requirements in terms of *minimum* size, with specimens in Caslon. We found, however, that, particularly with older children, a type-face might err by being too large as well as too small.[4] Hence we preferred to formulate our standards in terms of optimum sizes. The results [5] of our most recent experiments are shown in Table II. If a semi-bold type is used instead, the optimum size would be slightly smaller for ages 8 to 10 and slightly larger for ages above 12. We would add that, for the very

[1] For recent studies on this point, see M. Luckiesh and F. K. Moss, ' Boldness as a Factor in Type Design and Typography ', *J. Appl. Psychol.*, XXIV, 1940, pp. 170-83, and *id.*, ' Criteria of Readability ', *J. Exp. Psychol.*, XXVII, 1940, pp. 256-70. The writers experimented with Memphis type of four degrees of heaviness : but the differences found were small and in our view of doubtful significance.

[2] The term ' antique ' refers, not to the design, but to a particular kind of heavy face, distinguished from that called ' clarendon ' chiefly by the treatment of the serifs and by a wider set.

[3] If the reader will glance at (say) the prefaces to pocket Shakespeares published from Aldine House, he will see how surprisingly legible even the tinier founts can be. But these somewhat colourful faces are better suited to classical or poetic productions than to scientific work.

[4] Kerr points out that the largest of the large letters should in no way approach the limits of the young reader's area for distinct vision. In the British Association Report the largest of the specimen letters intended for beginners (who were apparently supposed to read letter by letter) was 0·16 in. high. Kerr observes that, at the ordinary reading distance, this would be about one quarter of the vertical size of the area of distinct vision. However, his estimate for this area appears to be derived from early experiments on adult vision, and greatly exaggerates the vertical span for the small child. But in any case, deductive determinations of this kind appear to us to be decidedly precarious ; and from more direct estimates we believe that the maximum has commonly been overrated.

[5] A rough notion of the effect of these requirements can be obtained by glancing at Burt's Graded Reading Test (Test 1, *Mental and Scholastic Tests*, pp. 340-1) which is intended for children aged 4 to 14 and was based on an earlier set of norms differing but slightly from those set out above. The same type face has also been used in the Standardized Attainment Tests recently issued by the Kent Education Committee (1954).

young (aged 9 or below), the words should be well spaced out—with an em-space as a minimum instead of an en-space as a maximum: and thin spaces should not be used in books intended for children.

TABLE II. TYPOGRAPHICAL STANDARDS FOR CHILDREN'S READING BOOKS

Age	Type-face (points)	Type-body (points)	x-height (inches)	Number of Letters in a Line of 4 in.	Length of Lines (inches)	Interlinear Space (inches)	Number of Lines per Vertical 5 in.
Under 7 years	24	30	0·150	30	5	0·250	12
7-8 years	18	20	0·111	38	4	0·170	18
8-9 years	16	18	0·090	45	3½	0·150	20
9-10 years	14	15-16	0·075	52	3¼	0·130	24
10-12 years	12	14	0·068	58	4	0·110	28
Over 12 years	11	12	0·060	60	4½	0·100	30

In Column 4, 'letters' means lower case characters or spaces, the space between two words being counted as one 'letter'.

For adults we used 8-, 9-, 10-, 11-, 12-, and 14-point Times Roman. With this face set solid, the most legible size was 10-point with an x-height of 0·062 in.[1] Judging by the average figures, any diminution in size (even by one point) and any great increase in size (two points or more) tend to reduce legibility. Although large letters are the most legible when read singly, they by no means favour quick reading or quick comprehension when used for consecutive prose: they cause visual attention to be directed to the letters themselves rather than to the form of the words, and to single words rather than to groups or phrases: the bigger the type, the smaller the amount of reading-matter falling within the normal eye-span, and the larger the number of eye-movements and fixation-pauses.

There were, however, marked differences in optimum size from one person to another, for the most part related to differences in visual acuity. With individual readers possessing normal eyesight the effects of small changes in size were insignificant. Older persons, even with apparently normal sight, often found 11-point more legible than 10-point. The majority of students (aged 19 to 26) found 9-point equally legible. But with presbyopic readers 9-point was apt to hinder comprehension and to induce eye-strain when reading was prolonged: these effects were evident, not only in the results of the questionnaire, but also in the irregularity of the eye-movements and in the increased blinking which became obvious to the investigator as he watched the reader.

(d) *Leading*. The introduction of 'leads' (i.e. thin metal strips inserted to widen spaces between the printed lines) greatly enhances the legibility of small type. In particular, it aids the eye to pick up the right lines as it moves back from the end of one line to the beginning of the next—a point of special importance with children who are extremely prone to doubling and skipping. In the table above, the 'interlinear space' is measured from the base-line, i.e. the bottom of the non-descending letters, to the top of the non-ascending letters of the line next below. Roughly speaking, the distance should be about half as wide again as the x-height—rather more with younger children, rather less with older—rather more with wide measures, rather less with narrow, in general, say about one-thirtieth of the measure. When the type used is large on its body or has long ascenders and descenders, it is essential that the leading should show a clear separation between the two lines especially when a descender comes directly above an ascender.

In this journal three sizes of Times Roman are used for the text—10-, 9-, and 8-point, all set solid. With these we found that the introduction of one or two points of leading appreciably increased the ease of reading. (With type-faces that are smaller on their body, like Bembo or Caslon, one point appeared sufficient.) Little seems to be gained by 3-point leading. 4-point leading usually diminished legibility. Like excessive letter-size, it tends to increase the number of eye-movements and fixation pauses. Plainly, the use of generous spacing can easily be overdone, as many readers have probably felt in perusing the first editions of the works of Ruskin or Tennyson.

[1] This is roughly equivalent to 11-point Baskerville and 12-point Bembo.

(e) *Measure.* Table II indicates what we consider to be the most suitable length of line for children of varying ages. The limiting condition is a matter not so much of visible inches but rather of the number of letters. With adults our experiments were in the main restricted to reading matter set in 10-point Times Roman. With this type, measures [1] shorter than 20 ems or longer than 33 ems diminished speed and ease of reading. Very roughly, these are limits of $3\frac{1}{2}$ to $5\frac{1}{2}$ in., 55 to 80 characters and spaces, or 2 to 3 lower case alphabets.[2] For literary material the narrower measure is desirable (say $3\frac{1}{2}$ to 4 in.); for scientific the wider. For a scientific journal intended for highly educated readers, who of course tend to skim rather than read word by word, a measure of 5 in. seems preferable.[3]

With long lines of solid type the eye finds it difficult to pick up the right line in turning from the end of a given line to the beginning of the next; and large pages filled with solid-looking panels of printed matter are apt to repel all but the hardened scholar. On the other hand, short measures, particularly when the type is big, prevent the eye of the trained reader from taking in large phrases with a single fixation and from making the most of the subsidiary help given in the horizontal direction by peripheral vision. Moreover, they necessarily entail widely varying spaces between the words, and increase the number of broken words at the end of the lines—features which appreciably hinder comfortable reading.[4]

Where the reading matter requires to be read mentally word by word (as in poetry), small measures, wide interlinear spacing, and even small type with fairly broad spaces between the words, are an advantage.[5] At the same time, small type further aids the eye to appreciate the shape of the stanza, especially where the metric structure is at all complex.

(f) *Margins.* There can be little doubt that books with excessively narrow margins are more apt to produce visual fatigue, though the diminished weight of the volume may then perhaps slightly lessen manual fatigue. Moreover, when the type-area extends nearly to the edge of the paper, the eye of the younger reader is apt to swing right off the page. With adults the effect of the broader margin would seem to be chiefly aesthetic.

To gain light on this problem we carried out a number of experiments upon readers' preferences in regard to the general imposition of the type upon the page. The results varied widely according to the subject matter of the book (literary or scientific), the size and shape of the page, the size, style, and interlinear spacing of the type, and so forth. To simplify the task we confined ourselves mainly to volumes whose pages showed proportions approximating to those of the golden section ($1 : \frac{1}{2}\sqrt{5}-1\frac{1}{2}$ i.e., 1·000 : 0·618) : this covers most of the ordinary octavo sizes. Expressing the margins as percentages of the measure, the weighted average of the preferred specimens gave the following proportions (rounded for convenience) : (i) inner margin 17·5, (ii) head 22·75, (iii) outer margin 27·5, (iv) foot 32·5. These are very nearly $\frac{7}{40}$, $\frac{9}{40}$, $\frac{11}{40}$, and $\frac{13}{40}$ respectively. The total of all four, it will be observed, adds up to the size of the measure.

Let us consider how this would work out if applied to a demy octavo publication (such as Spearman's *Abilities*, Thomson's *Factorial Ability*, Burt's *Factors of the Mind*, or the journal *Mind*). The dimensions of the trimmed page would be about $5\frac{1}{3} \times 8\frac{1}{2}$ in. ; and the foregoing recommendations

[1] The ' measure ' (i.e., width to which type is set on the printed page) is generally stated in terms of ' 12-point ems ', i.e., in multiples of one-sixth of an inch : (this unit is sometimes, but not quite correctly, designated a ' pica ' ; e.g., [22], p. 114).

[2] The last is a convenient mode of statement, since most specimen sheets give the length of the alphabets. With types of other sizes the same alphabetic rule will serve. If a large body is used (or the equivalent in lead), take the alphabet length of type corresponding to the body size: e.g., with 10-point type set with 2 point lead, take the alphabet length of the corresponding 12 point type.

In studies of legibility the influence of the width or ' set ' of a type face has been almost wholly ignored. And, although we ourselves have made no direct investigation of this aspect, we believe that its importance is by no means negligible, particularly in books for children. This indeed is one of the reasons why, from a psychological standpoint, we consider it preferable to state the length of the lines in terms of characters or words rather than in ems or inches : thus for the educated reader lines of 10 to 16 words—say 13 words on an average—seem most suitable.

[3] A line of 5 in. implies a page of about $7\frac{1}{2}$ in. wide, i.e., crown quarto. For books an octavo size is more convenient.

[4] On the basis of an admirably planned series of experiments Paterson and Tinker summarize the essential results as follows : ' One may characterize the oculo-motor patterns in reading an excessively short line by saying that the number of fixations is increased, the span of perception decreased, the mean duration of fixations increased, total perception time greatly increased, though the number of regressive movements remains approximately the same. . . . An excessively long line gives major difficulty in swinging back to the beginning of successive lines ' (D. G. Paterson and M. A. Tinker, ' Influence of Line Width on Eye Movements ', *J. Exp. Psychol.*, XXVII 1942, pp. 574-6). Their optimum range for 10 point type (80-88 mm. or 19 to 21 ems) implies rather narrower measures than that recommended here. But otherwise our findings seem perfectly consistent with theirs.

[5] One of us has an early edition of Tennyson's *Princess* in a long primer (roughly 10-point) with 8-point leading. Certainly the poet successfully compels his readers to read slowly !

would imply a measure of 22 ems or $3\frac{2}{3}$ in. and a depth of 38 ems or $6\frac{1}{3}$ in. (i.e. a type-area somewhat smaller than was actually adopted for the publications mentioned). With 10-point Times New Roman set on a 12-point body, each page would then contain 38 lines, i.e., about 2100 characters or 360 words. The margins would be: (i) backs, $\frac{2}{3}$ in., (ii) head $\frac{5}{6}$ in., (iii), fore-edge, 1 in., (iv) foot or 'tail', $1\frac{1}{3}$ in.[1]

Roughly speaking the two side margins should together take up a third of the entire width of the page, i.e., half the measure. When the interlinear spacing is increased, all the margins should be increased. For young children the proportions should be much wider, allowing plenty of room at the bottom for the thumb and at the sides for the untrained eye to swing to and fro. For infants the line should end with the end of a phrase, leaving the margins irregular.

For highly technical works a larger page may often be desirable, e.g., royal octavo ($9\frac{1}{2} \times 6$ in.), and somewhat narrower margins can be tolerated, and indeed, in the view of most readers, would even be preferable. For a technical journal a broader page is desirable, e.g., crown quarto (10×7 in., allowing a measure of 5 in.).[2]

The Interaction of these Factors. In the course of all these experiments it soon became obvious that the procedure adopted by most psychological investigators, namely, studying the effects of the four or five main variables separately—size, leading, line-length, style, and heaviness of face—is quite inconclusive, as indeed the practice of every good printer should suffice to show. For example, a line of 5 inches is easier to read than one of 3 inches provided 10-point or 11-point is used: when 8-point is used, the reverse is true. 9-point type, set solid in a measure of 5 inches, is harder to read than 10-point: but, set with 2-point leading, it is just as easy, at least for the average adult; and it is almost as easy in a measure of 3 inches if only 1-point leading is used. 3-point leading seems disadvantageous with ordinary faces; but with comparatively heavy type, and a face that is wide rather than condensed,[3] it is of definite assistance to children.

Practical Corollaries. In summarizing the foregoing results, it may be of interest to consider them more particularly in relation to journals such as those published by the British Psychological Society. It should again be emphasised that our conclusions are themselves merely tentative, and require further verification: other factors besides those we have studied here may conceivably have an overriding importance.

Were we guided by the data so far adduced, we should be tempted to infer that the length of line ($5\frac{1}{3}$ in.) which is adopted for the *British Journal of Psychology* (i.e., what until recently was called the 'General Section') is, if anything, rather too wide for the size of type employed; that adopted for the present journal (5 in.) is satisfactory for academic readers; that adopted for the *British Journal of Educational Psychology* ($4\frac{5}{8}$ in.) is well suited for a journal of a slightly more popular kind; while the short lines ($2\frac{2}{3}$ in.) recently introduced by the *British Journal of Medical Psychology*, which is set with two columns on a page, are likely to diminish speed of reading, though they may possess other advantages of their own. Much, however, must depend on the style of printing to which the reader is accustomed. And in books and periodicals dealing with advanced mathematics the more elaborate algebraic equations would of necessity be badly broken up if set in columns of $2\frac{2}{3}$ in.; indeed, it was for this reason that the editors of this journal at the very outset chose a rather wide measure.

[1] It was noteworthy that elderly readers (including the oldest of the present writers) strongly preferred a much larger head margin no doubt as a result of preferences acquired when such margins were more fashionable. This would incidentally keep the proportions of the printed panel, as well as those of the page, near to the golden section.

In works on book production it is often stated that diagonals uniting the corners of the page should also pass through the corners of the type area; that the inner margin should be half the outer; and that the proportions should be approximately $1\frac{1}{2}$, 2, 3, and 4. This mode of imposition satisfied few of our readers; and it was frequently complained that the inner margins were so narrow that, with a thick and tightly bound volume, the words nearest the inner edge were hard to read. It may be noted that the proportions suggested in the text imply that the printed area will be no more than about 50 per cent. of the total area of the page. The war-time economy regulations required that the type area should fill at least 58 per cent.

[2] The proportions preferred for the margins of a crown quarto page (the size of the Society's journals) were much the same as those cited above, with slightly narrower head and bottom margins. They fall midway between those adopted by the Cambridge Press for the *Brit. J. Psychol.* before and after 1942 (namely, 1 in., 1·4 in., 1·6 in., and 1·9 in. with a measure of 4·4 in., giving a type area of 30 sq. in., and 0·6 in., 0·9 in., 1·1 in., 1·2 in., with a measure of 5·3 in., giving a type area of 42 sq. in.). Note that in all these specifications the running headlines and folios are not counted as part of the type area.

[3] Times New Roman (327) has a rather narrow set: the width of the lower case roman alphabet, set in 8-point (x-height 0·052 in.) is only 105 points, whereas that of Perpetua 11-point, which has practically the same x-height, is 118 points. 11-point Imprint and Modern 7 (both with an x-height of 0·060 in.) have an alphabet width of 128 and 135. 10-point Times Roman with a larger x-height (0·062 in.) has an alphabet width of only 124: this condensation makes it, in our view, unsuited to children under the age of 10, and caused a number of our readers to deprecate its use for literary publications.

For mathematical work the Committee of the Royal Society [1] selected three type-faces—Imprint, Times New Roman, and Modern 7. For books we are inclined to favour 11-point Imprint (large face) on a 12-point body set in a measure of about 23 ems (3·83 in.), or 25 to 28 ems if there are numerous equations. For periodicals this face seems a little too large, while the 10-point seems a little too small ; and a larger measure would certainly be desirable. On the whole, the most favourable type for such publications appears to be 10-point Times Roman with long descenders and with 1-point leading : (2-point leading if the older version, with short descenders, is used). For subsidiary matter, 9-point with 1 point leading, which is almost equally legible, might be substituted. No serious difficulty seemed to be experienced with short footnotes of 8-point, even when set solid, particularly when they deal mainly with references or somewhat specialized comments. But a good many readers find it hard to cope with 8-point matter running to lines of 5 inches when continued for more than half a page.[2] On all these points we should welcome comments from our readers.

Introspections. In the course of all these experiments, especially when our examinees were adults, we tried to elicit comments and introspections. It was most instructive to observe how often the subjective impressions of the readers failed to correspond with their actual efficiency.

Certain readers, for example, reported that they found 12-point type more legible than 10-point, others that they found 9-point type quite as legible as 10-point type, when their actual performances demonstrated beyond question that they were wholly mistaken. Often they ascribed to the design of the type effects that were really due to the size of the type or to the leading. The spacing between the lines and the length of the measure produces illusory estimates of the relative size of the print : with both 9- and 10-point type, readers who found that the change from solid type to 2-point leading improved their accuracy or speed frequently explained that they found " the large type much easier " when the size of the face remain unaltered. Thus, as Mrs. Beatrice Warde has rightly observed, "What the book critic calls readability is not a synonym for what the optician calls legibility."[3] Nearly all tended to read with greater facility the kind of types that they preferred, and were inclined to confuse intrinsic legibility with their private aesthetic preferences. As we have seen, preference depends largely upon custom ; and throughout it seemed evident that *almost everyone reads most easily matter set up in the style and size to which they have become habituated.*

III. AESTHETIC PREFERENCES AND THE CLASSIFICATION OF TYPE FACES

The Study of Preferences. On looking at recent scientific publications both in this country and America, one is tempted to echo a remark of Bernard Shaw's : " It is a mistake to think that the modern author is insensitive to the beauty of a finely designed and well printed page : he positively hates it." No doubt war-time regulations, the increasing need for economy, and, in technical subjects, the fact that the better designed founts are deficient in ' mathematical sorts ' may be in part responsible. But authors and readers are themselves largely to blame. The scientific writer is apt to declare that " print is required for use not for ornament : its function is to be legible, not to look beautiful ". Scientific readers, so far as our results can be trusted, vary much more widely than others in typographical taste and interests—the least sensitive usually predominating. And both author and reader are quite content to leave such matters to the printer or publisher.

The best publishers and typographers are fully alive to the importance of psychological factors. They recognize, as Oliver Simon puts it, that " there are aesthetic considerations prompting people to prefer one type rather than another, which are closely linked with the emotions and difficult to define precisely " [21, p. 11]. In work on book production and in the journals of the trade there has of late been much discussion about the artistic qualities of different type-faces; and, from the earliest days of printing, sculptors, architects, painters, and engravers have applied their skill to the designing of type. Yet hardly any objective or systematic research has been undertaken on the psychological questions involved. The psychologist, though he has studied legibility and made sporadic investigations in almost every other branch of art both pure and applied, has remained singularly uninterested

[1] See *Notes on the Choice of Type Faces for Scientific Periodicals* (1950) drawn up under the Chairmanship of Dr. S. Zuckerman by the Consultative Committee for Co-operation with Printing Organizations.

[2] Quite independently, Mrs. Beatrice Warde (of the Monotype Corporation) suggested that the footnotes in the Journal might be divided into two columns. The Editor would like to express his indebtedness to Mrs. Warde for her kindness in responding to various questions about the printing of the Journal and for her valuable advice.

[3] *Monotype Recorder*, XXXII (1933), no. 1.

in the aesthetic aspects of printing—a form that silently confronts him every day of his life. The problems, moreover, are not without their practical importance. This is perhaps most obvious in the psychology of advertising. Here the readers' reactions, whether conscious or unconscious, towards the style of printing adopted by the advertiser may defeat the entire purpose of the advertisement. Yet once again, though numerous experiments have been carried out on visibility and display, hardly any attempt has been made to discover what particular qualities make different kinds of type-faces attractive or otherwise to different kinds of public.[1]

Methods. In planning a tentative inquiry into the more obvious problems, we have followed much the same general procedure as was used by one of us in previous investigations on personal taste in literature, art, and music. We collected examples of all the commonest type-faces now in regular use, and then submitted them to a number of persons of both sexes, with the request that they would arrange the various specimens in order of preference. Nothing was said about the differences being aesthetic. Indeed, it was part of our problem to discover how far, if at all, the judgments were comparable with other modes of aesthetic appreciation. As in our work with pictorial tests and with projective tests such as the Rorschach, we used two methods of factor analysis—'correlating persons' and 'correlating items'.

(a) *Correlations between Persons.* We began with what is commonly called 'P-technique', i.e. correlating the rankings submitted by the several individuals and then factorizing the correlation table by simple summation. The method was first tried with twenty persons selected to represent different classes of reader.

Their rankings revealed a moderately large 'general factor for persons', accounting for nearly 41 per cent. of the total variance. Such a factor implies the presence of a common tendency underlying the different orders; what is its precise nature we shall consider in a moment. The second factor was bipolar, and contributed 13 per cent. to the total variance. It subdivided the entire group of persons into two subgroups: their factor measurements indicated that, with few exceptions, those in one subgroup tended to prefer 'old' faces, while those in the other preferred 'modern'. However, a sample of only twenty persons is scarcely sufficient to establish such a classification with any finality. Several smaller factors were also discernible, each contributing about 8 to 12 per cent. to the total variance: for the most part they appeared to relate to such influences as age, eyesight, and habitual type of reading.

Introspections. All our subjects were asked to give reasons for their preferences. With typographical material the replies seem far less self-conscious—far less influenced by the natural desire to show up well, which so often impairs the sincerity of the responses to tests of pictorial taste or personality. Yet the content of the remarks is so illuminating that the whole procedure almost takes the character of a camouflaged 'projective test'. Only one of our critics had any technical knowledge of typographical differences; but many of them, particularly the book-lovers, seemed sensitive to what one of them called the 'atmosphere' set up by the type adopted.

The reasons offered could be readily classified into much the same psychological categories as were reached in our previous tests with other kinds of aesthetic material.[2] This will be evident from the extracts given in the appendix, but may be briefly illustrated here.

(a) *Subjective.* 1. *Associative.* (Of Modern 7 and Bodoni): "I have always disliked that kind of type because it reminds me of the print used in the French books we were compelled to read at Lausanne—the most hateful time of my life." (Of Caslon): "It seems to bring back memories of 'ye good olde days'—reading 18th century editions of Swift and Defoe in my father's library." (Of Perpetua italic): "It reminds me of inscriptions on tombstones or memorial tablets in a church." (Of Modern 7): "That for me is the only proper type for mathematical work, probably because I've always seen it in that kind of type." "With types it's much the same as with the place you live in or the author you are reading: they're all so much more interesting when you know something about their history."

[1] At the National Institute of Industrial Psychology, nearly twenty-five years ago, one of us took part in a research on the letter-press to be used in the advertisements of a well known firm of drug manufacturers. The differences in apparent effectiveness were unexpectedly wide. Still more striking are the results we have obtained on repeating the same experiment with a post war group. Though the data so far available are somewhat limited, it is difficult to avoid the conclusion that in the commercial world taste and fashion change far more rapidly and erratically than in the world of literature or journalism, and that on the whole they have markedly improved, while in the field of psychology they have definitely declined.

[2] See C. Burt [15], pp. 280-88; cf. also the classification of types based by Bullough and others on introspective comments explaining colour-preferences and the like: E. Bullough, *Brit. J. Psychol.*, II, 1907, pp. 111-23; C. W. Valentine, *The Experimental Psychology of Beauty* (1919).

2. *Emotional.* (Of Bodoni): "It seems to prick and dazzle. Horrible" (Of Caslon): "Restful and soothing." "Irritating! Those italics make me want to put them straight, like pictures hanging crooked on a wall." (Of Garamond italic with its different angles): "It makes me giddy to read it."

3. *Anthropomorphic.* "A type should have an individual personality: that's why these (Caslon and Garamond) go at the top." (Of Bodoni): "Cold, stiff, and rigid, like a row of black shirts on parade." "Intellectual." "Introverted." (Of Caslon): "Homely, unassuming." (Of Fournier): "Whimsical." (Of Baskerville): "Genteel." (Of Perpetua): "Aristocratic." (A general remark): "A type should have no personality of its own: you expect handwriting to reveal something of the writer's character, but a printing machine has no character to reveal. Now these types seem to be saying 'Look at me; aren't I ingenious [Baskerville caps]? Aren't I dignified [Bodoni]? Aren't I quaintly old-fashioned [Caslon]? Each is trying to express itself when it ought to be expressing the author's meaning."

(b) *Objective.* 1. *Intuitive.* (Of Bodoni): "It gives you a general impression of an orderly formal scheme—straight verticals on a straight horizontal, everything correctly fashioned and everything correctly finished." (Of Caslon): "I placed it first, not for any particular features that I can single out, but for its general look—its Gestalt-quality, if you like: it forms just the right kind of pattern made of just the right kind of shapes." (Of Perpetua): "Every letter strikes you at once as well-proportioned, like the columns, doors, and windows in an Adam façade."

2(a). *Rationalizations.* "I'm an arrant functionalist: to me a letter looks well when it's doing its job well—no frills or fancy touches." (Of Scotch Roman): "Letters to be printed with moveable types, instead of written by hand, should be all of the same size[1]; these come nearest of those you have shown me. On the other hand, in this passage (Times Roman) the S and the J are far too narrow; they should be the same breadth as the U; and the W is too broad—it would fit better if the strokes were crossed like that" (the Bodoni W). (A general remark): "In my view, the principle of parsimony requires that letters should be constructed of a minimum equipment of lines: three kinds of straight strokes—vertical, horizontal, oblique, and two kinds of circle—full size in O and Q, and half size in S and B, with both halves exactly equal. Evidently, then, letters like S and B should be exactly half the width of O. N should be a perfect square; and letters life E and F exactly half the width of N. E should be two small squares—cross strokes all equal.[2] This (Centaur) comes closest to my own ideal. The more they violate these principles, the lower the types go in my list. The proper way to judge them would be to measure their dimensions under the microscope."

Most of our readers sought to judge all type faces by the same fundamental criterion, differing from one reader to another. There were, however, a few exceptions. Several of the more critical asked if they could alter their rank according to the subject matter for which the type was to be used. Four insisted that it was impossible to suppose that one and the same rank-order would hold for every kind of publication. 'Old' faces and ornamental details (like the swash capitals in Baskerville italic, or the unusual k, z, and Q in Garamond) appealed to two readers as suitable for poetry or historical romances, but were pronounced quite inappropriate for a daily paper or a scientific journal. Three remarked that they liked the reprint of a classic to be set in a style suggestive of its period. One put the principle still more explicitly: "I should like a plain machine-made face for scientific books and articles, a slightly more artistic style for what they call *belles-lettres*, a decorative type for imaginative prose, and a slightly archaic form for the older writers or stories of bygone days. . . . I should hate to see Pater or Malory set up in this type (Bodoni), or Kelly's *Statistics* in either of those (Perpetua and Garamond). There's nothing intrinsically bad about any of them. Aptness is my only criterion."

Nevertheless, an explicit recognition of these different functions was comparatively rare, even among our more literary readers. It was, indeed, surprising to note how few appeared in any degree typographically conscious. Many started straightaway with some deprecatory remark, to the effect that "all kinds of type look much the same to me". Yet, after a few moments' scrutiny, about one in three would become either wholesale critics or ardent admirers of this or that set of features in some particular group of faces.

(b) *Correlations between Type Faces.* To investigate the nature of these differential preferences we compiled a special set of test-material consisting of extracts cut from printers'

[1] The idea that every printed letter should fit into a rectangle of the same size and shape was a principle tacitly or explicitly adopted by many of our readers. Here possibly the influence of the typewriter may be discerned.

[2] It was surprising to discover that even our psychological critics were quite unfamiliar with the curious optical illusion whereby the top halves of the S, 8, etc., appear much larger than they are (as may be readily seen on turning the page upside down). If, therefore (as this reader demands), the letter S was mechanically constructed so that the top curve had exactly the same size as the lower, the top curve would, curiously enough, *look* very much larger.

specimen books, keeping so far as possible to the same or similar prose passages for all type-faces, but including in every instance a complete alphabet, in upper and lower case, both roman and italic. We selected type with an x-height of 0·062 in. or thereabouts (10-point for the larger faces, e.g., Times New Roman, 11-point for the smaller, e.g. Imprint and Modern 7).[1] These were submitted to 93 judges—teachers, students, University lecturers, and educated readers of the commercial or industrial classes. Most of them also took part in short tests of pictorial and of literary appreciation : (those described by Dewar and Williams [17, 18] were used).

A group of nearly a hundred is too large for P-technique. We therefore decided to ' correlate tests ' rather than persons. For this purpose we first converted the rank orders into normal deviates ; these were then reduced to deviations about the average for each type-face. Finally, we calculated all the inter-correlations between the different type faces, and factorized the table so obtained. It shows two peculiarities which have occurred—though not with any frequency—in other investigations on personal preferences (e.g., in a study of pictorial preferences by Sybil Crane) : (i) the largest factor is not a general factor with positive saturations throughout, but bipolar[2]; (ii) after the general and first bipolar have been removed, the presence of residual submatrices with figures approximating to zero indicates the presence not of further bipolars, but of group factors. The final set of factor saturations was obtained in the usual way by arithmetical rotation. The results are set out in Table III.

In many respects, as will be seen, though not in all, the classification indicated by the pattern of signs resembles the broad, semi-historical classification adopted by most books on typography. To secure further light on the reasons for the divergences and similarities, we endeavoured to secure introspections, particularly from typical representatives of the more dogmatic critics. Relying mainly on these two lines of approach—quantitative and qualitative—we may provisionally put forward the following interpretation of the more conspicuous factors.

I. There is first a small general factor which here accounts for only 3 per cent. of the variance. More than half its saturations are devoid of statistical significance. Evidently the factor simply indicates that the more general differences between the type-faces were not completely abolished by taking an ordinary unweighted average when the rank-orders reduced to deviations.

II. The largest factor of all is a bipolar factor marking the contrast between (A) *Old Face* types and (B) so-called *Modern Faces*. Two of the specimens in our series—Baskerville and Times New Roman—had saturations which, though positive, were statistically non-significant.

III. After the general and first bipolar factors had been removed, the table of residuals, when appropriately arranged, showed four square submatrices containing fairly large figures, while the rest of the figures were nearly all non-significant. To fit this pattern four broad group factors were obviously required. Of these (A1) the first classifies together a set of ' old ' type faces modelled chiefly on Italian or French designs : (two smaller group factors distinguish the two). I shall call this group Continental old faces. (A2) The second includes all the British old faces. The other two factors subdivide the ' modern ' faces. And once again the main division is into (B1) Continental and (B2) British. Those who are interested in the details of this classification will find them discussed in the Appendix.

Implications of the Factorial Results. We do not suggest that these statistical results make any new contribution to what we may call historical or comparative typography.[3] Their implications are of psychological rather than typographical interest. Supplemented by the introspective comments and biological details obtained from our various readers, the data as a whole throw considerable light on the psychological nature and the apparent causes of the varying responses which different type faces excited. There are two main questions.

[1] This decision eliminates one factor that appeared in our first exploratory studies. Certain sizes are undoubtedly more favourable to certain type-faces. Thus, Imprint, Plantin, and Times New Roman seem most effective in medium sizes : Caslon, Garamond, and Centaur show up best in larger sizes. Since this is itself merely a preliminary inquiry, several important conditions had to be ignored. Perhaps the most important is that of paper and ink. Half-tone blocks, such as are required for many scientific works, need glossy paper; and on these types like Caslon or Garamond would be quite unsatisfactory, whereas Bodoni looks well.

[2] For a discussion of correlation matrices in which the largest factor is a bipolar, not a general, factor, see P. Slater and C. Burt, this *Journal*, IV, 1951, pp. 9-20.

[3] In theory the saturations should indicate which particular type face is most representative of the family to which it belongs. But the data are too slender for much importance to be attached to these minor differences.

TABLE III. SATURATION COEFFICIENTS FOR TYPE FACES

TYPES	FACTORS							
	General	Old-Mod.	Cont.	Brit.	Ital.	Fr.	Cont.	Brit.
A. Old Face								
1. Continental								
a. Italian								
Bembo	+·173	**+·746**	**+·342**	−·062	**+·433**	−·063	−·084	−·034
Veronese	+·106	**+·318**	+·182	+·165	**+·226**	+·037	−·033	+·093
Centaur	+·127	**+·443**	**+·217**	+·144	**+·271**	−·042	+·096	+·116
b. French								
Garamond	**+·230**	**+·732**	**+·304**	−·058	−·024	**+·446**	−·028	−·045
Granjon	**+·204**	**+·728**	**+·295**	+·095	−·052	**+·424**	+·106	−·069
Fournier	+·191	**+·581**	+·342	−·107	+·067	**+·519**	−·090	−·167
Plantin	**+·212**	**+·465**	**+·239**	+·151	+·105	**+·350**	+·152	+·132
Ehrhardt	+·095	**+·403**	**+·378**	−·083	−·043	**+·285**	−·047	−·144
2. British								
Caslon	+·148	**+·714**	−·113	**+·536**	+·040	−·031	−·079	+·030
Baskerville	+·173	+·170	+·145	**+·443**	−·067	+·073	+·181	+·078
Imprint	+·184	**+·429**	−·132	**+·515**	−·031	−·045	−·075	−·065
Old Style	+·165	**+·357**	+·107	**+·369**	+·071	+·062	+·036	−·081
Times New Roman	+·081	+·187	−·083	**+·270**	+·056	−·058	−·041	+·121
B. Modern Face								
1. Continental								
Bodoni	**+·219**	**−·781**	+·179	−·121	+·033	+·054	**+·534**	+·059
Didot	+·175	**−·660**	−·123	−·032	−·051	+·167	**+·373**	−·084
Walbaum	+·161	**−·636**	+·160	−·045	+·046	−·053	**+·367**	−·022
2. British								
Bell	+·102	**−·324**	−·107	+·118	+·054	−·092	−·058	**+·480**
Scotch Roman	**+·346**	**−·542**	+·131	−·064	+·012	+·129	−·102	**+·435**
Modern 7	**+·364**	**−·715**	−·096	−·086	−·038	+·151	+·085	**+·323**
Contrib. to Variance (per cent.)	2·7	30·4	6·8	4·2	1·9	5·1	3·4	3·5

Figures in Bold Type are statistically significant.

First, what determines the *general* order of the type faces—their aesthetic quality, their legibility, or perhaps some other typographical characteristic? Secondly, what determines the *individual* differences—aesthetic appreciation, visual acuity, or perhaps some kind of temperamental characteristic?

General Order. The first factor obtained with P-technique may be regarded as a general factor for 'typographical taste'. The factor saturations for persons can therefore be correlated with the factor measurements for 'general aesthetic taste', based on the supplementary tests of pictorial and literary appreciation. The correlation is moderately high, 0·63. This provides a provisional answer to one of our questions. So far as these results go, we may fairly conclude that the differences between type faces on which these preferential judgements are based are largely, though not perhaps wholly, *aesthetic* differences.

Orders for Main Group and Subgroups. With P-technique the factor measurements for test-items give a weighted order of aesthetic merit for the type faces used. It will, however, be more instructive to examine the orders obtained in our second set of experiments, since these were based on a far larger sample of persons.

In Table IV we give the averages of the individual rankings for both roman and italic founts. In addition it seemed advisable to separate our readers into two main groups according to their interests or previous education: for convenience we may call them the 'literary' and the 'scientific' groups respectively. The rankings have been averaged for each group separately; and it will be seen that they reveal suggestive differences. In the table the various type faces have been arranged according to the average preference for *roman*

founts, since this would seem to provide the best all-purposes order. The weighted order given by the general factor for persons is almost identical with the average ranking given by the ' literary ' group.

TABLE IV. PREFERENCES FOR DIFFERENT TYPE FACES

Type	Roman			Italic			Combined Total
	Literary group	Scientific group	Total	Literary group	Scientific group	Total	
Imprint (101)	1	6	1	3	5	3	1
Times New Roman (327)	7	2	2	8	7	6	2
Bembo (270)	3	9	3	2	10	4	4
Plantin (110)	2	13	4	7	6	5	5
Baskerville (169)	11	5	5	11	4	9	7
Modern (7)	15	1	6	4	1	1	3
Caslon (128)	6	12	7	6	8	8	8
Bell (341)	13	4	8	10	3	7	9
Scotch Roman (46)	16	3	9	5	2	2	6
Granjon (Linotype)	8	14	10	13	16	14	11
Centaur (252)	4	18	11	1	13	10	10
Veronese (59)	5	19	12	12	12	13	13
Garamond (156)	9	17	13	9	11	11	12
Fournier (185)	10	16	14	15	15	15	15
Old Style (2)	14	11	15	16	9	12	14
Ehrhardt (453)	12	15	16	14	19	16	16
Didot (520)	17	8	17	17	18	19	17
Walbaum (374)	18	7	18	18	17	18	18
Bodoni (135)	19	10	19	19	14	17	19

The Bipolar Factor. The second factor obtained with P-technique, it will be remembered, was a bipolar factor, distinguishing those who prefer old faces from those who prefer modern ; and, in accordance with the so-called ' reciprocity principle ', the first bipolar factor obtained with R-technique is, here as elsewhere, virtually the same as the first bipolar factor already obtained with P-technique. The former—those who (according to the factor measurements) preferred old faces—turn out to be, with few exceptions, students or lecturers in the Faculty of arts ; the latter—those who preferred modern faces—include most of the regular readers of scientific or mathematical works, a small proportion of the ' literary group ', and the large majority of the non-academic (i.e., commercial or industrial) readers. In these cases, therefore, the individual deviations from the general order of preference appear to be in part at least an effect of habituation ; and the presence of this large bipolar factor implies, or at any rate is consistent with, the view that different kinds of type are suited to different kinds of job.

There were, however, a number of sufficiently striking exceptions to indicate that familiarity and habit are by no means the only influences. Once again, the results of our supplementary tests are instructive. They indicate that a preference for the older type faces is associated with what I have loosely called ' romantic ' tendencies in other forms of artistic appreciation, while the preference for modern type faces is associated with a ' classical ' taste [1] : there is in fact a correlation of ·43 between the two sets of factor measurements.

Legibility and Preferences. For any given type face and any given person it is possible, by the methods described in the foregoing sections, to derive two distinct measures—(i) for legibility or rather ease of reading (obtained by combining the assessments for speed and accuracy) and (ii) for aesthetic preference (obtained by converting ranks into standard

[1] Cf. [17], p. 275, [18], pp. 298 f. In these earlier researches evidence was obtained which indicated that ' romantic ' and ' classical ' preferences were in some degree dependent on temperamental differences. Even if this were also true of typographical preferences (and on this we had no direct information), it would not follow that the typographical preferences were *directly* affected by temperamental differences : the differences, for example, might conceivably have influenced the choice of academic subjects.

measure). Thus for each separate fount we can compute a coefficient of correlation showing how far, if at all, these reactions tend to go together. For each of the roman types in our list the correlations were positive and (with one exception) significant: they ranged from ·23 for Times New Roman to ·47 for Modern 7.

To secure further evidence on the significance of this relation, Mr. Cooper constructed a Snellen test-chart with Modern 7 letters, and tested the visual acuity of each available reader at a distance of 6 metres. The correlations for visual acuity were ·68 with ease of reading and ·36 with preference. Eliminating the effects of differing visual acuity by the usual statistical procedure, we obtained a partial correlation between ease of reading and preference amounting to ·33—fully significant with the number tested (53). Taken in conjunction with introspective evidence and other observational data, the result strongly supports the view that in practice legibility is not merely a matter of the size and shape of the black marks on the white paper or of the physiological efficiency of the eye: there is also what may be called (in the fashionable phrase) a 'psychosomatic' influence at work: printed matter seems more legible, and reading becomes more accurate and quick, when the material is set in a type which the reader, perhaps without realizing it, finds aesthetically pleasing. The reader, of course, rationalizing his impressions, puts it the other way round.

Supplementary Factors. As we have seen, the remaining factors group the type-faces into families and indicate how far their special peculiarities harmonize with one another, and influence, in varying degrees, the preferences of individual readers. It would seem that, just as the style of type evolved by different nations or by different epochs reflects their cultural characteristics, so the preferences of our readers are themselves related, often in a somewhat complex and indirect fashion, to their more specific interests and cultural attitudes. Such influences are particularly obvious in the case of those outstanding individuals who in the past have sought to make a contribution to the theory or practice of typography. How far it is true of ordinary individuals further research alone can prove.

Implications for Psychology and Education. It would scarcely be appropriate here to enlarge on the corollaries that might be drawn for the psychologist or for the educationist and teacher. Two points, however, deserve brief mention. On comparing the results obtained in our more recent experiments with those obtained by one of us some thirty years ago, we have been struck by the marked changes in typographic preferences that have taken place during the intervening period; and this has brought home to us the need, and the possibility, of extending such investigations in the temporal or time-dimension by using what the factorist would probably call 'O-technique'. Indeed, a psychological study of the gradual development of taste and style during the four hundred years that have elapsed from the time of Gutenberg and Jenson to the present day might well provide a simple and instructive contribution to what has been called, perhaps a little ambitiously, the 'psychology of history'.

At the same time we venture to suggest that those who hold that "a more prominent place in the curricula for boys and girls should be assigned to aesthetic subjects and developing a capacity for the appreciation of art and industrial design" [1] might legitimately include in their efforts the appreciation of good craftsmanship in typography, in printing, and in the production of books. "A good piece of lettering" (it has been said) "is as beautiful a thing to see as any sculpture or painted picture." [2] And the work we have observed in some of the L.C.C. Schools of Arts and Crafts shows what can be achieved in this direction.[3]

IV. SUMMARY AND CONCLUSIONS

1. By means of tests of accuracy and speed, supplemented by observations of eye-movements, blinking, and other symptoms of eye-strain, an attempt has been made to determine the relative legibility of different styles of printing. The chief novelties in the experiment were the inclusion of newer type-faces issued during recent years and the application of the crucial tests after a period of preliminary reading to permit fatigue or adaptation.

2. It was found that the different characteristics—size, design, boldness of type, width of measure, of margins, and of interlinear spacing—all influence and interact with each other,

[1] Report of the Consultative Committee of the Board of Education on *Secondary Education*, p. 171.
[2] Eric Gill, *An Essay on Typography*, p. 22.
[3] It may be added that the first edition of Burt's *Causes and Treatment of Backwardness* was set up, printed, and bound by boys in one of the schools of the National Children's Home.

so that assessments obtained by varying just one characteristic in isolation may at times be highly misleading. Revised norms for reading books have been computed for children of different ages, together with an indication of the most suitable measures and leading. On the whole, Old Style Antique appeared most appropriate for children under 12, and Imprint, Plantin, or Times New Roman for those over 12. With adult readers enjoying normal vision wide variations in design, size, or measure seemed permissible without greatly affecting efficiency of reading. Type having an x-height of about 0·060 in. (e.g. 10-point Times New Roman or 11-point Imprint or Modern 7), with 2-point leading (or 1-point with type small on its body) and a measure of 20 to 33 ems, proved to be most satisfactory for general purposes. Slightly wider measures and somewhat narrower margins seemed preferable for technical as contrasted with literary or popular publications.

3. The application of the methods already used by psychologists in research on other forms of artistic appreciation—ranking of items and a factorial study of the results—furnished an independent classification of type-faces similar to what may be called the historical classification, but differing suggestively in minor details. This psychological classification of type faces implies a corresponding classification of persons according to their distinctive preferences ; and the factor measurements showed significant correlations with the type of aesthetic appreciation found with other test-material. Partial correlations indicate that apparent legibility may itself be strongly influenced by half-unconscious preferences.

4. The introspective data obtained during our experiments on typographical preferences disclose a highly complex motivation—the customary reading and the cultural interests of the reader playing an unexpectedly important role. The reasons offered by our subjects in explaining their likes or dislikes reveal much the same motivational types as have been noted by Bullough and others in their work on the appreciation of colours, pictures, and musical chords ; and the content of the responses often showed a close resemblance to those obtained from projective tests with inkblots or geometrical designs.

5. The conclusions reached in this preliminary report must be regarded as provisional only. The study of preferences for different styles of printing appears to be an unduly neglected branch of the psychology of industrial design, possessing considerable practical importance for the printer and advertiser and much theoretical interest for the psychologist. Numerous incidental problems were encountered in the course of our experiments that call urgently for more intensive research.

APPENDIX

THE CHARACTERISTICS OF THE COMMONER TYPE FACES : READERS' COMMENTS AND HISTORICAL NOTES

Modern Developments in Printing. As reported in our previous issue, the Council of the British Psychological Society has decided to refer the choice of type-faces for its various publications to the new Publications Committee. If, however, such discussions are to be fruitful, it is plainly essential that those concerned should know what type faces are now available and what are their specific characteristics.[1] We know of no short publication giving this information in a compact and non-

[1] There is, of course, a prior issue. Many members of the Society hold that, in order to preserve continuity, there should be *no* change in the type or format of any journal. Others believe that there should be one type and format for *all* publications of the same society. It will be observed that on both these issues we ourselves take the opposite view. We may note that, in response to our appeal for the views of readers and members, several have complained that (if we may quote one correspondent) ' in spite of the striking advances made in British printing during the last 20 or 30 years, most of the Society's journals are still printed in the same type and style that they adopted before the first world war '. The medical journal, it may be noted, changed the type-face and the printed area in the middle of Vol XIX (1943), though the editor states that he desired to make the changes 10 years earlier. *Psychometrika* has quite recently changed its style of printing.

We ourselves are particularly indebted to the generosity of the Monotype Corporation, of Linotype, Ltd., of Intertype, Ltd., and of Messrs. Stephenson Blake (Caslon Letter Foundry) for presenting us with specimen books, specimen sheets, etc., usually available only to the trade. Examples showing the more distinctive letter-forms in the various type faces that we used will be found on p. 57. The terminology used in describing them is based on that proposed by Mr. Joseph Thorp [12]. We have also to thank Mr. David Thomas (Typographical Consultant to the Publications Department of University College) who was good enough to read these pages in proof, and suggest several important corrections.

technical form. The same want has been felt in discussing the selection of type for children's books and for verbal tests ; and several teachers, psychologists, and education officials have asked for some such synopsis. A detailed enumeration of distinctive peculiarities will have little interest for the ordinary reader ; but it is indispensable for our purpose, since (as we have frequently observed) an unsuitable eccentricity in one or two letters is apt to be overlooked until the type face has been chosen and the material actually set up. In citing examples, we have, so far as possible, tried to name books and periodicals which would be familiar to psychological and educational readers, and at the same time illustrate the use of the italic as well as the roman forms. Those who have taken part in our experiments and discussions have frequently inquired about the origins of the various designs, and of the names by which they are known : and consequently we have included a few brief notes on the sources of the commoner faces. Indeed, without some knowledge of their history, and of the conditions under which they were produced, it is difficult to appreciate the nature and appropriate functions of the different types.

Origins. Printing is an art in which, for both historical and technical reasons, the personal element and the mechanical are blended as they are in no other. This double character, as may be seen from the introspections of our subjects, is largely responsible for the conflicting criteria to which they variously appeal, and for the unique and complex problems in aesthetic psychology which the study of printing presents. The production of books was the earliest industry to which methods of mechanical standardization and mass production were applied.[1] The idea of using movable types was apparently due to Dutch woodcutters. The special invention of Johann Gensfleisch (' Gooseflesh ', known as Gutenberg, from the village from which his family came) was the punching of metal matrices and the use of adjustable moulds in which movable types could be cast in large numbers. For the results to be intelligible to his readers the printer had necessarily to imitate the handwritten letters with which they were already familiar. Gutenberg (1454), like Caxton after him (1474), printed in ' Gothic '. What today we call ' roman type ' was evolved in Italy, between 1465 and 1470, from a curious combination of Roman inscriptional majuscules, dating from the period of Trajan's column (2nd century, A.D.), and a Renaissance revival of the Carlovingian minuscules.[2] As a result the shapes of our modern capitals are derived, not from the attractive curved uncials, developed by scribes during the fourth and fifth centuries, but from square geometrical forms incised on stone with a chisel ; on the other hand, the smaller characters of our ' lower case ' still preserve much of the qualities of handwritten letters, notably the oblique thickening produced by the slit and slightly canted nib of a reed or quill.[3]

A. *Old Faces.* The types commonly grouped under the phrase ' Old Face ' are characterized (i) by light ' colour ' (in the printer's sense), (ii) by comparatively slight differences between thick and thin strokes, (iii) by bracketed and sloping serifs, (iv) by the oblique or ' biased ' shading which is characteristic of pen work, and (v) by figures which do not range on the line.[4]

As noted above, those of our readers who obtained positive factor measurements for Bipolar II (mainly readers in our ' literary group ') preferred types belonging to this family. It was here that the conflict between the two rival criteria became most evident. This is clearly discernible in some of the introspective comments quoted above. The following are still more pertinent. " This style of printing (Modern 7) seems lifeless and machine-made ; the others (Bembo and Caslon) look as though they had been drawn by a human hand." " Those types (Bodoni and Scotch Roman) look too geometrical ; a book is a work of literature—of art ; and the type should have something of the organic grace of a freehand line." With these contrast the criticisms offered by the other group. " That

[1] Many of our older readers, especially those who had been influenced by the teaching of Morris and his followers, were inclined to protest against what they supposed to be ' the gross deterioration in printing produced by the application of mechanical methods '. But the printing press is itself essentially a machine ; and such protests are five centuries too late. Shaw's secretary relates an instructive story. " G. B. S. wanted the type (for his collected edition) to be set up by hand : in the Morris tradition, he scorned machinery and the usurping monotype." The managing director of R. & R. Clark arrived with two specimen pages, and invited him to pick that which he thought the best. Shaw unhesitatingly chose the page which (as it turned out) had been set by monotype ! So far from debasing the quality of book production, the introduction of mechanical type-setting has greatly enhanced it : and there can be little doubt that what has been called ' the revival of printing ' and the present prestige of British typography have been largely a product of the courageous policy of the Monotype Corporation, who, 50 years ago, embarked on a progressive recutting of classical type-faces, supplemented by new designs by artists like Bruce Rogers and Eric Gill [23].

[2] During the revival of learning initiated by Charlemagne after the dark ages (*c.* 790), Alcuin of York was invited to the Abbey of St. Martin at Tours to superintend the copying of Church books. The story goes that Charlemagne, who himself desired to write, declared that " a hand accustomed to the sword could form only the simplest shapes " ; hence the substitution of the plain half-uncial forms for the confused and multiform Merovingian hands. Being without ligatures, this simplified style lent itself admirably to the requirements of movable types when the time arrived.

[3] Those readers who wish to examine further the typography of the printers named in the following paragraphs will find original specimens of the work in the cases in the King's Library of the British Museum which illustrate the History of Printing.

[4] Some ' old faces ' are now supplied with an alternative set of figures ranging on the line.

sort of type (Garamond and Fournier) may have been quite all right when people wore picturesque clothes and lived in Renaissance palaces : in a technological age you want a plain technological type." " This kind of lettering (Baskerville) sets me wondering who drew those pretty curves, when I ought to be thinking who wrote this passage and what did he mean." " Modern printing should be like modern building, purely mechanical and functional, not a phoney revival like late Victorian Gothic or pseudo-Tudor house-fronts."

It will be remembered that the factorial analysis of the specimens in our series subdivided the Old Faces into two main sub-groups, called for purposes of reference ' Continental ' and ' British ' ; and the former in turn was re-divided into what we shall call ' Italian ' and ' French ' respectively.

1. *Continental.* The most distinctive features of the ' Continental ' subgroup are the crossed W and the splayed M. Both these features were criticized by teachers and others who thought them " too eccentric " for children's books or " too archaic " for scientific books. " The legs of the M seem to be slipping apart ", says one. " Nowadays ", says another, " double-U is no longer a double letter, and V no longer stands for U." Others, however, thought these minor variations " add life and interest to the printing ".

(a) *Italian.*[1] There is no single criterion distinguishing all the specimens in this subdivision from the next. To judge by the comments of our readers, the common characteristic is largely negative, namely, a freedom from the decorative tendencies that are so marked in the other type-faces in our continental series, particularly the French. The nearest approach to a positive diagnostic clue is the curve of the *v* and *w* in the italic founts—neither angular nor rounded, but (to quote one reader) " a bit *nouvel art* ". However, so far as it can be defined, what seems to have led most of our readers to rank all three specimens together at much the same level, either high or low according to their taste, is an impressionistic quality difficult to describe, but summed up by one of them (a little sarcastically) as " a chaste art-and-craft appearance ".

Bembo (270).[2] This type, a comparatively recent revival, is perhaps the best representative of the early classical design from which all our later roman types are descended. It won high praise from both our ' literary ' and our ' scientific ' groups. A tiny point which raised it (with Plantin) above all the other ' Franco-Italian ' types was the lower case j, with its generous curve and bulb, which prevents any confusion with i. The x-height is small, to leave room for the long ascenders and descenders ; indeed, the ascenders are taller than the capitals—a feature occasionally criticized : (as one reader remarked, " the uneven heights of the first two letters in ' The ' or ' When ' suggest that the printer had unintentionally picked then from different founts "). As to the italic (which comes from a different source, see p. 54) opinions were divided. Teachers who were members or admirers of the ' Society for Italic Handwriting ' praised the oblique serifs and the sharply curved terminals in the lower case, and thought the use of this type face in reading books might be associated with a revival of the so-called Italic style for children's script. On the other hand, non-academic readers often considered these features " rather mannered ". The italic *y* (" more like a Greek gamma with a serif to its tail ") was almost universally condemned.

Veronese (59). This was one of the earliest of the artistic designs cut by the Monotype Corporation, and was brought to them in 1911 by J. M. Dent (publisher of the ' Temple Classics ' and ' Everyman's

[1] The letters in the earliest printed documents (emanating from Gutenberg's firm, *c.* 1454) were based on a degenerate vernacular (' Burgundian Bastardas '—full of sharp points like Gothic architecture) which had become current North of the Alps and was derisively dubbed ' Gothic ' by Vasari. As ' Schwabacher ' and later ' Fraktur ' it was proudly preserved throughout the centuries as the German national style. Elsewhere the change to so-called roman letters was due to certain historical accidents.

After driving out the English, Charles VII (the king crowned at Rheims during the days of Joan of Arc) sent Nicholas Jenson to Mainz to study the new invention. When Gutenberg's foundry was burnt in the sack of Mainz, two of his assistants fled to Subiaco near Rome, and there installed a printing press, the first to be set up in Italy, at the Monastery of Saint Scholastica, later transporting it to the Palazzo Massimi in Rome. Jenson apparently went with them; but he eventually moved on to Venice (1470). Since the books were intended for Italian scholars, it was natural to adopt the old *littera antiqua* or *humanistica* then in favour with the leading humanists. Jenson's designs were based on the best manuscripts, and are the most highly praised of all classical faces: (for reproductions see [5] and [25]). On his death his material passed to the father-in-law of Aldus Manutius. The name ' roman ' apparently originated in France, and was used to distinguish the Italian type-faces both from the ' gothic ' or black letter and from Aldine ' italic ' : it seems to refer, not to Jenson's adoption of monumental Roman capitals for his majuscules, but to the style adopted by the printers at Rome.

[2] In 1495, soon after setting up his press at Venice with material from Jenson, Aldus Manutius printed a small tract, *De Aetna*, written by the 25-year-old Venetian poet Pietro Bembo, one of the leading humanists, famous for his revival of Ciceronian Latin, and later Cardinal and Secretary to Leo X. The type, cut by Griffo of Bologna, differed somewhat from the designs of Jenson : the e's and h's have their more familiar shape ; the capitals are narrower, and their slab serifs abolished. Aldus himself seems to have discarded this type almost at once in favour of yet another version, cut for his still more celebrated *Hypnerotomachia Poliphili* : it was this that apparently furnished the basis for Garamond's French types, and so ultimately for Caslon and his successors.

Those who wish to judge the suitability of Bembo for psychological work will find good instances of it in Ryle's *Concept of Mind* (Hutchinson, 1949) and Valentine's *Parents and Children* (Methuen, 1953).

Library '). It was modelled, partly no doubt as a result of William Morris's influence, on a rather heavy Venetian 15th century original.[1] As several of our readers observed, it is suited only to a restricted range of publications. Many teachers, particularly headmistresses, thought it excellent " for children's poetry books " or " for Shakespeare ". The points that most frequently aroused comment were the slab serifs of the M and N, the sloping stroke of the e (both reversions to Jenson's design), and the thickening of the horizontal strokes in the z. The italic is a modified version of that associated with the Italian and French type-faces, rounded, expanded, and slightly thickened to secure greater legibility.

Centaur (252). This type-face was originally designed by the American typographer, Bruce Rogers, for Houghton Mifflin.[2] The roman fount is modelled on Jenson's roman, with triangular serifs for the lower case, but without Jenson's slab serifs for the capitals. The most distinctive features are the sloping cross-stroke for the lower case e, the incurved shank of the italic *h* (both a reversion to Jenson's designs), the absence of thickening in the tail of the *j*, and the straight thin stroke of the *y*, thickened at the lower end. In the re-drawing, the type seems to have lost much of the robustness of Jenson ; and was even criticized by some of our readers as " self-conscious " and " art-y ".[3] The italic, designed by Frederick Warde, is modelled on that of Arrighi (see p. 54).

(b) *French*.[4] The two peculiarities most frequently criticized in this group were the pointed bases of italic *v* and *w* and the swash descending tails of the italic *z* : most of our readers thought such letters highly unsuitable for algebraic symbols, and many disliked them even for general use.

Garamond (156). The roman resembles the Venetian characters of Aldo, but the j preserves the old form, that of an elongated terminal i : the similarity caused a good deal of unfavourable comment. But the chief target for criticism was the italic—" sometimes too narrow, sometimes too wide, and the capitals sloping at inconsistent angles ", reflecting " an unskilled stage in the development of type-design ". (Compare the slanting *Q* and *S* with the nearly upright *R*, the splayed *A*, *V*, and *x* with the narrow *S*, *B*, and *y*.) The *Q*, " shaped like a written 2 ", the curly *k*, the long-tailed *z*, the sharp-based *v* and *w*, the *g* with its wide lower bowl, the *y* with its back-swept tail, and the primitive hooped *h*, " almost like a *b* ", were disliked by many. A few warmly praised these peculiarities, as " giving the type a marked character of its own ".[5]

Granjon (Linotype).[6] We included this because we were told that it was one of the most popular linotype faces. The design has many similarities with the preceding : e.g., the j tapering to a point. It differs chiefly in using the same form for the capital J (almost a straight line), avoiding the crossed W, and reverting for the italic *v* and *w* to the form used in Bembo (the last two features probably account for its being ranked above Garamond). Its most distinctive peculiarity is the long downward tail to the italic *k*—criticized by all who disliked any tendency to flourish.

[1] A similar type, based mainly on Jenson's roman, had been designed by Emery Walker for the short-lived but celebrated Doves Press (1901) at Hammersmith. It should be noted that much of William Morris's typographical work was due to the skill and learning of Emery Walker. Morris (according to his daughter) after hearing a lantern lecture by Walker on ancient type-faces, suggested that they should join in founding the Kelmscott Press (1890). Emery Walker later became Chairman of the Consultative Committee on Printing Schools under the L.C.C. (1920).

[2] Bruce Rogers joined Emery Walker for a while at Hammersmith and later served for a time as adviser to the Cambridge Press.

[3] The psychologist will find an interesting example of this type—italic as well as roman—in Sherrington's *Man on his Nature* (Cambridge University Press, 1940).

[4] The essential characteristics of what are now called ' old faces ' became fairly well fixed during the later years of François I—' the golden age of French typography '—largely as a result of his passion for Italian art and culture. Claude Garamond (*c.* 1535), the most successful of the French type-cutters of the period, produced what eventially became the standard European type outside Germany. It seems certain that copies of Aldus's work reached Garamond in Paris ; and, as we have seen, it was on these that Garamond's designs were based.

In the 17th century, after a temporary decline, French printing, like French scholarship, revived under Richelieu ; and the Imprimerie Royale, established by him in 1640, started with what were taken to be Garamond types. Thus, the style that had been introduced a century earlier continued to dominate French printing, with minor variations, until it was superseded by a ' modern face ' at the end of the 18th century. A ' Garamond ' type was recut in France in 1898, and it was one of the first of the classical faces to be cut by Monotype. Mrs. Beatrice Warde (' Paul Beaujon '), however, has shown that the punches acquired by the Imprimerie Royale in the 17th century were not, as they supposed, those cut by Garamond, but were in fact designed, nearly a hundred years later, by Jean Jannon of Sedan in 1621 [8].

[5] Professor Saw's *Leibniz* (Penguin series, 1954) is set in Garamond. It was, we think, an unfortunate type to choose for Kenyon's *The Bible and Archaeology*, where the juxtaposition of the italic *g* and *y* (as in *Egypt* and *archaeology*) requires a very ugly logotype.

[6] Granjon was printer to Henri II (*c.* 1557) and afterwards to Pope Gregory XIII. The type, produced by G. W. Jones for Linotype, was based on 16th century French type faces. Mention should perhaps be made of another fount recently cut by the same designer, and called Estienne (after a celebrated ' Royal Printer ' of the same period). It has the old-fashioned italic seen in Garamond, but it is chiefly noticeable for its long ascenders and descenders—a feature revived by Eric Gill in his ' Perpetua '. Estienne and several other faces we were forced eventually to exclude for fear of overloading our tests.

Fournier (185).[1] Several readers were puzzled by the serif to the foot of the roman b (seen also in Bodoni) ; and the lower case j (with no blob) was again the subject of criticism. Many also thought the roman fount too condensed, especially the M and Z : (the 12-point lower case alphabet has a length of only 130 points, i.e. 1·8 in., as compared with Garamond which has 144 points, i.e. 2 in.). The italic evoked the most opposite views. Unlike the preceding founts, instead of providing a new set of designs, it is to a large extent merely a sloping version of the roman : thus the *m* and *n* have sharply sloping serifs instead of curved initial hooks, and the upper right arm of the *k* has no loop. Several considered the inclination excessive (19° instead of the Garamond 12°) ; and the oblique bowl of the *g* and long downward tail of the *z* were frequently condemned as " irritating " or " trying too obviously to be quaint ".

Plantin (110).[2] This type face, designed in 1913 by the Monotype Corporation, was the first produced for use on coated paper.[3] Its blackness on uncoated paper has made it popular for children's books. The italic, with the well rounded base to *v* and *w*, departs widely from the French designs, and was thought to be well adapted for mathematical use. Nevertheless, rather to our surprise, it was far from popular with our scientific readers—chiefly because of the roman j (" too like i "), the crossed W's, and the italic *J* and *Q*. One or two noted the gap in the P between the upright and the end of the curve, and asked if it was deliberate or accidental ; and several disliked the flat top to the capital A.

Ehrhardt.[4] The roman j tapers to a point ; both arms of the C have serifs ; the M is splayed ; the italic *J* has an unpleasantly low bar ; and in both founts the Q has an elongated tail—features often prompting unfavourable comment. Several also disliked the somewhat irregular bowls of the roman and italic g's. The characteristics most frequently condemned, however, were the peculiar italic *w* (joined head-strokes, pointed bases, while that of the *v* is round) and the even line of the zero—" more like a circle than a numeral ". This was the least popular of the old faces.

2. *British*. England, unlike Germany and Italy, had no national pride in printing, and readily borrowed from the other side of the Channel. Caxton's first book (*The Game and Playe of Chess*, 1475) was printed at Bruges in an unattractive Gothic style ; and one of his later types (based on the formal liturgical hands—*textura* [5]) became the basis of the so-called ' English black letter '. Shortly after the establishment of the Royal Society, Dr. Fell,[6] Bishop of Oxford, imported punches from Holland, which then had almost a monopoly of type-founding, and with these (which still exist) set up the Oxford Foundry. The first competent engraver of types in England was William Caslon, whose foundry has survived to the present day.

The type-faces classed together in the ' British ' group by our factorial results are distinguished by several obvious points from the French and Italian. The roman letters, particularly the capitals, have a wide set. The W's are uncrossed, and the M's unsplayed ; the roman Q (except in recent forms) has a swirled tail. The j has a well rounded terminal ending in a bulb. The italic comes closer to the English running hand, and is somewhat condensed. The Dutch origin is traceable in the high bar and swirled head to the italic *J*, and in the curved serifs of the rounded italic *v* and *w*.

Caslon. Caslon started a new foundry in 1720, modelling his letters on those of van Dijck, but greatly improving their delicacy. The result was a distinctly English face, which (except for a

[1] The Fourniers managed the Le Bé factory in the 18th century. P. S. Fournier (' le jeune ') proposed the ' point ' system in his *Manuel* (1764), and started modifying the old French designs in the direction of the *gout hollandais*.

Bertrand Russell's *Inquiry into Meaning and Truth* (Allen and Unwin, 1940) and Murdo Mackenzie's *Contrast Psychology* (Allen and Unwin, 1952) are set up in 12-point Fournier type.

[2] Towards the close of the 16th century, superiority in printing, as in other forms of scholarship, passed for a time from Roman Catholic France to Protestant Holland. Christopher Plantin (1514-89), born near Tours, established his press at Antwerp, and shares with Christoffel van Dijck (type-cutter for the Elzevirs) the credit of producing some of the finest Dutch work. Their designs were sturdier versions of those produced by Garamond and Granjon. There is an Intertype Plantin with more irregular italic. Neither version is by any means an exact reproduction of the original.

[3] It stands up well to the demands of high speed rotary printing. The most familiar instance of its use is *The Listener*. For the psychologist an instructive example is the new edition of Trotter's *Instincts of the Herd* (Oxford University Press, 1953), which shows the blackness of the type on uncoated paper.

[4] The type appears to have been suggested by that used by Ehrhardt of Leipzig (*c.* 1720) and derived by him from the Dutch foundry of Anton Janson.

[5] The so-called *lettre de forme*, a large upright hand, with sharp angles and no curves, having a woven ' texture ' : (Cf. ' Fair as a text B in a copy book,' *Love's Labour's Lost*, v, ii, 42). By the 15th century it had superseded the fine hand of the Winchester School (11-12th century, a blend of Anglo-Irish with Carolingian). The Authorized Version (1611) was printed in *textura*. Caxton himself explained that his books were " not for a rude vplondyssh man to rede, but onely for a clerke and gentylman ", and he therefore aimed at " a meane bytwene the playn and curyous " (Prologue to *Eneydos*).

[6] His name survives in popular memory from the doggerel parody of Martial's 33*rd Epigram*, written by Tom Brown as an undergraduate, after Fell (then Dean of Christ Church) had expelled him. Besides Caxton's press in the precincts of Westminster Abbey, presses had existed at Oxford and St. Alban's from the 15th century. Of the several early foreign printers who set up presses in London, Richard Pynson, afterwards printer to Henry VIII, used a ' white letter ' to print an *Oratio* for the papal nuncio (1509).

spell of neglect in the 19th century) has remained in continuous use for over 200 years.[1] Perhaps because the Declaration of Independence was printed in Caslon, it has been still more popular in the U.S.A. The minor irregularities of shape and weight give it, as our readers put it, a " homely, unassuming, old world look " ; and led others to condemn it as " unskilled " or " primitive ".

In the roman the distinctive features are the wide H, M and U—the H indeed being frequently criticized as " disproportionately broad ". Several disliked the obliquely flattened top to the A, and asked if the type was not slightly battered. The irregularities are most marked in the italic. In both Caslon and Baskerville the italic *Y* resembles a Greek upsilon (a shape which excited repeated criticism) ; the lower case *w* has an open loop ; and the C in both roman and italic has a well marked serif to the lower arm (sometimes omitted in smaller sizes of Caslon roman). With the Caslon italic the downstrokes to the angular letters, *A*, *V*, and *W* are parallel to those of the square letters, *H*, *M*, *I*, *K*, etc., so that the former seem to be toppling forward. As a result, the gap between the capital initial and the rest of the letters sometimes breaks the word in two—a point seized on by most of the teachers.[2]

Baskerville.[3] The most conspicuous peculiarity of Baskerville is the pointed unserifed junction of the two inner strokes of the large and small W—a feature commended by many of our readers, who argued that " in English W is a single sound, not two consonantal V's ". In the g the lower bowl is not completely closed. And practically all the letters display an increased contrast between the thin lines and the thick.

Owing to a free use of kerned letters the italic has an individuality of its own. The terminals are rounded, and the thickening has shifted from the bases to the down strokes—characteristics suggestive of elegant penwork. Compared with the roman, the italic shows even greater condensation than Garamond, and was consequently criticized as " crowded ". Considering the popularity of this type with many publishers, the criticisms we received were unexpectedly frequent. Some thought the general effect " a little too genteel ". The decorative tendencies of the italic excited both praise and blame. The " pothooks and hangers " of the italic *m* and *n* and the upstroke to the italic *p* were said to be " suggestive of copperplate engraving " ; the curves of the *k* (as in Fournier devoid of a loop) and of the *z* were often considered " fanciful ", and the swash italic *K*, *N*, *T*, and *Z*, and the generous tail of the Q's (roman as well as italic) " much too distracting ".[4] On the other hand, several readers placed this type at the top of their list ; and many picked out the very characteristics that others so strongly criticized as " making the type far more interesting than ordinary humdrum print ".[5]

Old Style. The design that originally bore this name (cut in 1860 by Phemester for a well known Victorian firm, Miller and Richard) sought " to supply a type with old face qualities without the archaic features of Caslon ".[6] It was so successful that it was quickly copied by most other founders in this country and by printers in Germany, Scandinavia, and the U.S.A. Thus, by the beginning of the twentieth century the majority of books were printed either in a variant of Old Style or in one of the ' Moderns ' : and probably these publications moulded the notions of our older readers as to what the proper shapes for English letters should be. There are numerous minor variations of the ' Old Style ' face : e.g., Monotype and Intertype have an A with a pointed apex, while Linotype, like the earlier founts for handsetting, usually has a flat top. We used the Monotype version (Series 2)—one of their " early bread and butter founts " (as they call them) cut in 1901. The serif on the upper limb of C and G has a well marked slope, and the G and b have spurs. The absence of hairlines and the even colour were praised by most of our readers. Comparatively few *specific* points were criticized ; but a number of our younger literary readers thought the general appearance " rather insipid ", or " harmless but lifeless ".

[1] Its renewed popularity at the beginning of the present century was due partly to Bernard Shaw. Shaw supervised all his printing : (cf. his paper ' On Modern Composition ', *Caxton Mag.*, 1902). The plays had to be set in Caslon long primer unleaded—" no rules, no flowers, no headings, side or cross ". But how many have been repelled by the solid grey oblong pages of his prefaces, often without a single indentation!

[2] For an example of Caslon see G. F. Stout, *God and Nature* (Cambridge University Press, 1952).

[3] John Baskerville (b. 1706) was a writing-master at Birmingham. After seven or eight years spent on designing a type suitable for " books of consequence ", he became printer to the University of Cambridge. His smooth wove paper and novel printing techniques made it possible to produce a far clearer and more delicate impression ; and, as Macaulay puts it, his *Virgil* " went forth to astonish all the librarians of Europe ".

[4] Monotype provide their Caslon fount with alternative italics possessing similar flourishes, though (except for titling) they are rarely used.

[5] For examples of Baskerville, see Ayer's *Language, Truth and Logic* (Gollancz, 1946) and Ferguson's *Young Delinquent in his Social Setting* (Oxford University Press, 1952), where the non-ranging numerals and the swash italic capitals in the tables seem open to criticism. Baskerville has of late been an exceedingly popular type for ordinary novels.

[6] Caslon was revived in 1844 by the Chiswick Press for a fictitious 17th century *Diary*. The old Fell types were resurrected in the following year, and used to print Tupper's *Proverbial Philosophy* and the *Oxford Book of English Verse*. In the last quarter of the century Ruskin, Emery Walker, William Morris, and Charles Ricketts initiated " an aesthetic typographical renaissance " that eventually made " modern English printing the soundest in the world " [5]. See above, p. 48, footnote 1.

Imprint. This was " an all purpose design " cut by the Monotype Corporation in 1912 to the specification of G. T. Meynell and J. H. Masson for a new monthly review called *The Imprint*. It proved to be one of the most popular faces in our series. The general style is that of Caslon ; but the type is larger on its body, and gets rid of the peculiarities that strike the modern eye as primitive—e.g., the marked and irregular slope of Caslon italic, the serif to the lower arm of C, and the curved italic *Y*. The roundness of the lower case italic resembles that of Baskerville, but the central loop to the italic *w* is not visibly open. The one feature most frequently criticized was the tiny tail to the roman Q.[1]

Times New Roman. The Times newspaper, originally set in a rather crude old face, changed early to modern (1799), and kept it for over a century. On October 3, 1932, the paper arrived at the breakfast table in an entirely new type, specially designed by Mr. Stanley Morison (Reader in Bibliography in the University of Cambridge and Typographical Adviser to the Cambridge University Press and the Monotype Corporation). A year later the design was released for general use. At the present day it is probably more widely used than any other face. For its original purpose, to combine legibility with economy of space, it is admirably adapted—large on its body, compressed in spite of its x-height, yet with well open counters, and furnished in its original form with short space-saving ascenders and descenders. We found it amazingly legible in the smaller sizes.

Its more obvious characteristics classify it with the ' old ' faces. The serifs to the capitals are bracketed ; while those of the lower case are wedge-shaped ; the shading of the lower case roman is strongly biased (indeed, the tops of the curves in b and p and the bottoms in c, e and q were sometimes criticized as " disproportionately thick "). C and G are clearly distinguished; the J is non-descending; the Q has a small central claw tail (perhaps leaving the letter " too much like an O "). The italic commence with straight serifs, and the *v* and *w* have angular bases. Mathematical readers objected that some of the italic was unsuitable for algebraic equations, in particular the *v*, which is indistinguishable from a Greek *nu*. On the whole, however, this face was one of the very few that were popular with both the literary and the scientific groups.[2]

Perpetua (239).[3] Because of its exceptional interest, we included this type face in our first and larger series ; but, since it has rarely been used as a book face,[4] we dropped it in our final experiments. The roman *looks* an extremely legible face ; but many readers found that after a while it proved rather tiring. The italic aroused the most conflicting judgments, even among our literary group. There was, however, a general agreement that, for text matter as distinct from a title page,[5] its peculiarities tend to thrust themselves between the author and the reader. Accordingly, in view of its rather distracting quality and the unusual form of the numerals, we had little hesitation in deciding that it was quite unsuitable for mathematical publications.

B. *Modern Faces.*[6] The distinctive characteristics of the so-called modern faces are (i) the vertical instead of biased shading, (ii) the marked contrast between the thickness of the ' down strokes '

[1] For examples of Imprint the reader may take McDougall's *Psychology* (Oxford University Press, 1952, originally set in Scotch Roman, 1912) or Burt's *Subnormal Mind* (Oxford University Press, 1935).

[2] This Journal and the *Brit. J. Med. Psychol.* are set in Times New Roman.

[3] Perpetua (called after an African martyr) first appeared in an exhibition of types arranged at the St. Bride Institute in 1930 ; but it seems to have attracted little interest from the printing trade until after the war. It was based on drawings made by Eric Gill, *not* (so he says) with special reference to typography. The roman follows the style of the best chiselled inscriptions of the Roman era, with Gill's own stone-carving as an intermediary stage. The narrow E and S, the slightly splayed M, the generous tail of the Q's, the thrust-out leg of the R's, and indeed the proportions generally, seem manifestly copied from the celebrated panel on Trajan's Column. The U, a letter which does not exist in Latin, is uncial. Elsewhere Gill declares that the designs preserve " that commonplaceness and normality which is essential to a good book type ". But much of the italic is neither commonplace nor normal. With both founts the idiosyncrasies are best seen in the larger sizes. The stress is usually vertical ; the serifs are hairline but bracketed, horizontal in the roman, oblique in the italic ; the O is circular ; the E equal-armed ; the bottom of the d and the top of the q are flat ; the ear of the g is horizontal ; the upper stroke of the 5 unusually long. The italic is a slanting roman. The *B*, *D*, *R*, and *P* have flourished headstrokes, and the *U* has both its verticals thickened. But the most striking peculiarity is the loop of the *g*, which turns up and joins the bowl, so that the letter resembles an inverted *B*.

[4] It has recently been used for the text of several Penguins (e.g. *Selected Poems*, by T. S. Eliot, 1952) ; and the more recent linotype version, called Pilgrim, has been adopted for the current issue of *The Yearbook of Education* (Evans, 1955). To the ordinary reader the style will probably be most familiar from its use for headlines in newspapers and advertisements catering for educated or sophisticated readers (e.g. those of the Arts Council, H.M. Stationery Office, and even the wrappings of household commodities intended for the ' more refined classes ').

[5] For titling and display there was a general agreement among our testees that Perpetua was " the most delightful of any style of type ". Of the type-faces suggested for the cover of this *Journal* it received by far the largest number of votes.

[6] The so called ' modern ' faces became popular in France at the close of the 18th century when ' classical ' (i.e., Roman) art became fashionable at Paris. The alleged model was the rigid sculptured lettering—the *capitalis quadrata*—of the ancient inscriptions, many of them newly discovered, which were deemed aesthetically superior to the freer and more flowing characters traced by the monastic quill. The

(i.e., those that are vertical or run down from left to right) and the thinness of the curved strokes and the ' upstrokes ' (i.e., those that run up from left to right), (iii) the fine horizontal unbracketed serifs (iv) the figures ranging on the line, and (v) generally, the geometrical and rigidly mechanical structure of the design.

As we have noted in our studies of other forms of art, there seems to be a kind of ' temperament ' (if we may use such a word to designate what is quite as much a cultural, epochal, and often national attitude as an effect of innate constitution) which tends to admire geometrical, symmetrical, and perpendicular qualities in human art and to favour a scientific or technological design as contrasted with what is imaginative or emotional or gives free rein to fancy—in short, a classical taste rather than a romantic. During periods in which one or other of these tendencies becomes dominant in sculpture, architecture and literature, it generally appears, after a little delay, in commercial and domestic art, and ultimately in typography. The correlation between the various manifestations was obvious not only in the results of our tests, but also in the introspections. People who showed the preferences of the ' stable introvert ' in other forms of aesthetic appreciation (cf. 18, p. 297) seemed also to reveal them in their preferences for type faces ; and frequently they used the same arguments to justify both forms.

Here, as before, the factorial results subdivided the various type faces belonging to this broad class into two subgroups, the Continental and the British.

1. *Continental.* In the ' Continental ' subgroup the characteristics distinctive of the ' modern ' face appear most clearly. The ' British ' as usual show an inclination to compromise.

Bodoni (135).[1] Morris vehemently objected to " the sweltering hideousness of the Bodoni letter, the most illegible ever cut, with its preposterous thicks and thins ". Our readers generally ranked it last of all. Numerous incongruities were noted. Since the alphabet includes letters, like V, W, X, and Y, which are composed solely or mainly of symmetrically arranged oblique strokes, it is impossible to apply the principle of vertical thickening with complete consistency. Moreover, in the upper case the inner limbs of the W cross, in the lower they meet in a point, as in Baskerville. The C has a serif to the lower limb, like Baskerville ; the b a foot serif, like Fournier. The top of the t is cut off flat. The italic is only slightly sloped (10°) : some letters begin with a horizontal serif like printed roman ; the *v* and *w* (but not the *u*) with a rounded curve like a running script. The right hand upper limb of the italic *y* is curved outwards ; and in both roman and italic the capital J falls only a little below the line—" as though the printer could not make up his mind ". The squareness of certain letters, the correct symmetry of others, the parallelism and rigidity of the lines, led many readers to complain of the ' geometrical ' design : " it might almost have been drawn on squared paper with a ruler and compasses, and then carefully shaded in." On the other hand, several of our subjects observed that, owing to its emphatic character, it seemed highly suited (in the larger sizes) for commercial advertising or headlines in newspapers.

result was a somewhat stiff and pompous style, like that of academic painting and sculpture during the years when Louis David was director of arts. It should be noted that an upright style of writing existed long ago in this country (e.g., that used for the Lindisfarne Gospels, an insular adaptation of the half uncial hand). The alphabet then in vogue contained no v, w, or y : with these letters, in order to avoid making *both* downstrokes thick, the method of writing has to be changed ; and so an obvious incongruity arises.

Many teachers who favour an upright style of writing require their pupils to hold the pen at right-angles to the ruled or imaginary base line on which they write. The upright characters of the early manuscripts, however, were obtained, not by changing the position of the hand, but by cutting the nib obliquely.

The common notion that the ' modern ' style was first introduced by Bodoni and then taken up by the Didots is quite mistaken. Similar characteristics are traceable in several Italian manuscripts of the late renaissance. Italian architects and painters, and later the French savants of the Académie des Sciences under Louis XIV, had constantly interested themselves in inscribing letters in squares, triangles, and circles according to definite proportions, and in attempting to formulate artistic taste in terms of mathematical principles. As we ourselves discovered, it was a popular hobby with several male teachers and craftsmen who took our tests : see, for example, the ' general remark ' made by one of our readers, and quoted above under ' rationalizations ' (p. 40).

[1] Giambattista Bodoni (b. Piedmont, 1740) was head of the Stampa Reale at Parma. The rather dazzling brilliance of the contrast between the heavy main strokes and the finer hair strokes, and possibly the fact that such type required good paper and was itself a sign of an amazing technical precision, rendered the new style popular for all serious work. Bodoni printed editions for Gray and Horace Walpole, who warmly applauded the ' new continental workmanship '. He started with French models. Indeed, the thin horizontal unbracketed serifs that are so characteristic of the ' modern face ', as well as the narrowing of the letters, are already descernible in type cut by Grandjean (1745), who had been ordered by Louis XIV to produce a new design that should be the monopoly of the Louvre—the Romain du Roi. But it seems to have been the new delicacy and emphasis introduced by Baskerville which Bodoni and the Didots chiefly admired and endeavoured to magnify. The new Bodoni Press in Italy has recently produced remarkable examples of the typographer's skill. In this country, however, it has become confined to titling, advertisement, and letter press accompanying art books on ultra smooth paper.

The versions of ' Bodoni ' cut by British and American foundries are more open than the original, and the heart shaped italic *v* designed by Bodoni has been discarded (except by German printers) for a shape conforming with the *w*.

Didot (520).[1] This type face is intermediate between the older French faces and Bodoni. The italic is almost identical in form with the roman, and its inclination is slight. The wide *H* and *U* and the pointed *j* were freely criticized.

Walbaum (374). This is a German 'modern' type (1810), revived by the Curwen Press and later by Monotype. The design is based on Didot; and both roman and italic are wide and therefore readily legible. The roman, with its non-descending J, spurred b, e with a large counter, and square-topped t, suggests a lighter version of Scotch Roman, though the pointed j indicates its French origin; the italic *v* and *w*, though wider, are reminiscent of Bembo and the double curved *y* of Bodoni. The descenders of the *p* and *q* have no serifs. Walbaum and Didot were among the least popular faces.[2]

2. *British.* English printing has always been characteristically light. Hence in Britain, even modern book faces have tended to avoid the excessive thickening of their continental prototypes, though in the 19th century Thorne and his imitators set a fashion for 'fattened' styles that have remained popular with commercial firms. For the most part anything glaringly foreign was discarded, and several distinctively English features due to Caslon and Baskerville were preserved. At the same time the British versions introduce one or two points that are wholly new.

Bell. What is sometimes called the earliest English modern face was cut in 1788 by Austin for the London bookseller and publisher, John Bell, who was the first to discard the long-tailed s. In contrast to the condensed modern faces of Continental printers, its well-rounded letters give it a distinctly British look. It was recut by Monotype from the original punches found by Mr. Stanley Morison in a lumber-room of Stephenson Blake (the Caslon Letter Foundry). Like other modern faces, it has horizontal serifs, perpendicular shading, and a large counter to the roman e. It has a straight serif to the italic *j*, and a plain italic *J* standing on the line (though the roman drops below). On the other hand, it retains the peculiar italic *Y*, with the curved upper arms adopted by Caslon and Baskerville, and reverts to the excessively sloped *V* and *W* of Caslon. The lower limb of the italic *k*, *K* and *R* are curved, like those of Baskerville; and similar curves now appear in the roman k, K, and R. The italic *x* is also curved—formed by two *c*'s placed back to back. These curvilinear innovations were widely criticized by our readers; but with the scientific group it was one of the most popular of the older founts.[3]

Scotch Roman (46). In matters of literature and the arts Scotland has always kept in close contact with France; and there has been, and still is, some degree of trade rivalry between the printers of Edinburgh and Glasgow and the presses of Oxford, Cambridge, and London. It is not surprising, therefore, to find that the changes distinctive of the new French printing at the close of the 18th century were quickly reflected in Scottish publications. In 1815, Alexander Wilson, a Glasgow printer, introduced what he called 'Scotch modern face'—a sturdier version of Didot's types. The ascenders and descenders are shorter, and the serifs have slight brackets. The most distinctive features are the non-descending J, the Q with the tail beginning with a little circle inside the counter, and the curly lower limb to the R ending with an upward turn like an italic *m* or *n*. As in Bodoni the top of the lower case t is square. The italic is not unlike that of Baskerville, but less condensed and more emphatically shaded. The *p*, however, begins with a hook like *i*, *j*, *m* and *n*; and the *x* resembles that of Bell. The y is new, with a round bowl like that of the ordinary running hand. Indeed, many of our readers complained that the italic resembled the pseudo-copperplate writing of the cheap printed visiting card, " with 'thin upstrokes and thick downstrokes', as the old fashioned schoolmaster used to say". This became one of the most popular founts of the 19th century.

Modern Series 7. A 'modern' type face (Series 1) was cut by Monotype as early as 1900. Series 7 (an 'extended' version) includes over 600 'mathematical sorts,' and is the fount adopted by the authors of *The Printing of Mathematics* (Oxford University Press, 1954). It is thus a specially important type face for our purpose. In its most obvious characteristics it resembles Scotch Roman. Most of the serifs, however, are unbracketed; but the thin verticals have brackets: thus U has a hair-line at the top of the left hand upright and a triangular serif at the top of the right hand—an inconsistency which several readers noted. Unlike Scotch Roman the roman t has a pointed top as in most other type faces. In the italic, *g* has a plain oval bowl, like a written g, and a tail like the *y*; the *w* has a plain central downstroke instead of a loop; the *v* and *w* start with a simple hook (like *m*, *n*, and *u*), not with the circumflex curve of older British founts.

Among our scientific readers Modern 7 appeared the most popular of all. The non-scientific criticized it sometimes on patriotic grounds, sometimes on aesthetic: many objected that it had "an un-English" or a "French" look; others thought it "too rigid and mechanical"; several

[1] In France the *gout nouveau* was developed by the Didot family at the very beginning of the 19th century: the most celebrated were printers to Monsieur (afterwards Louis XVIII) and the Comte d'Artois (afterwards Charles VIII). The elder, F. A. Didot, himself a noted classical scholar, reformed the point system, making the unit equal to $\frac{1}{72}$ of the pre-metric inch. Their design, still more condensed, has remained the standard style for French printing, where, during the last 150 years, typography has changed far less than in this country.

[2] Walbaum is used for Grey Walter's *The Living Brain* (Duckworth, 1953).

[3] *English Social Differences*, by Professor Pear (Allen and Unwin, 1955) is printed in 11 point Bell.

who laid claim to typographical taste believed " it was time that scientific publications attempted a style of printing less crude " (or " less inartistic " or " less old fashioned ", i.e., 19th century).[1]

Italic. In general italic type differs so widely from roman that we considered it essential to obtain preferential orders for the two founts separately.

What we now call italic was first introduced by the Venetian printer Aldus Manutius in 1501 for his pocket editions of the classics. For a more popular kind of reader it was natural to turn to a more popular kind of letter—the everyday cursive which had grown up side by side with the formal. Moreover, a running hand tends inevitably to slope ; and the slope allows greater condensation and so economizes space : (with Baskerville the italic takes only 87 per cent. of the space required by the roman). Griffo, who cut the type, is said to have taken Petrarch's writing for his model.[2] These early attempts were somewhat crude and found few imitators : in particular much difficulty was experienced in sloping the roman capitals so as to give them the appropriate inclination. The standard italic is derived not from Aldo but from Vicentino (Ludovico degli Arrighi of Vicenza), a writing master in the Vatican Chancery, who (like Baskerville later on) turned printer and published (1522) a copybook claiming to teach *littera cursiva o cancellerescha* : this was a neo-Carolingian minuscule, used for papal ' briefs ' (i.e., short notes and informal correspondence) ; and nearly all the ' swash ' capitals that appear in later faces are to be found among Arrighi's decorative majuscules. Robert Estienne,[4] one of the most celebrated scholars and printers in the reign of Henri II, adopted Arrighi's letters as his model ; and most later italic founts are descended from that. ' Italic ' remained for long an independent rival to the ' roman ' letter. Being less bound by convention, it has varied far more widely. Hence it is always easier to recognize a type-cutter from his italic than from his roman.

In England ' the sweet Roman hand ' (*Twelfth Night*, III, iv, 31) only became popular with the cultured classes about the time of Elizabeth I. Half a century later, as England gradually acquired the carrying trade of the world, English commercial clerks developed a more rapid and rounded script, which became known abroad as *anglaise coulée*. The letters of Snell's commercial copybook of 1711 differ but little from those of Vere Foster on which the oldest of the present authors was brought up. The English type-cutters of the 18th century naturally modified the Franco-Dutch italic to conform with the national style. Baskerville, who was himself a writing master, and the later printers of ' modern ' types, tended more and more to follow the flowing script of the copperplate engravers. During the 19th century, when italic ceased to be an independent type-face, one would have expected it to harmonize more closely with its roman context. Early attempts in this direction, as we have seen, were made by Fournier and others at the close of the 18th century ; but in this country the principle has still secured only a half-hearted acceptance.

The motives which seem thus to have determined all these different styles are not unlike the reasons given by our readers to explain their own preferences. Those who prefer a scholarly and dignified hand admire the humanistic pattern of the older Italian type-cutters ; those whose own handwriting tends to flourishes enjoy Baskerville with its swash capitals ; those who look for a hand that is swift, practical and free from stylized mannerisms prefer the smooth and uniform roundness of Modern 7. Many of our teachers in particular stressed the fact that children often try to " write like a book ", and produced several instances in which pupils suddenly altered their lettering on the introduction of a book with a novel type face. (A familiar instance is the occurrence of the Greek *e*, *d*, and even *a* in the handwriting of classical scholars.)

Among readers who took part in our experiments the sharpest clash of opinion arose between those who like their italic to contrast with the roman and those who think it should conform and harmonize. But here again much obviously depends on the reader's implicit notion of the functions of such a type—a point too frequently ignored. Is it to be used for prefaces, for passages quoted from ancient authors, for initial abstracts or summaries, for mathematical theorems, to provide an algebraic notation, to emphasize a conclusion or key phrase in the text, or merely for side-headings, stage-directions, and the like? For emphasis or headings many thought it more logical to use, not a cursive and therefore a lighter type, but heavy semi-bold roman. But all this raises yet another of the many problems awaiting psychological research. The layout of commercial advertisements has of late been rendered far more successful by the use of trial inquiries planned by a statistical

[1] In our preliminary experiments we also included ' Century '—one of the modern type-faces that has apparently achieved considerable popularity in America. It was rated quite highly by a small group of readers, who happened to be familiar with its use in a certain newspaper, but low by literary readers. The extremely short descenders and the large counters (notably in the e) no doubt are particularly suitable for newspaper work ; but were criticised by many of our subjects, especially the teachers.

[2] A comparison with Petrarch's holograph does not bear out this pleasing tradition : see also Updike [5], I, p. 128.

[3] See A. F. Johnson and S. Morison, ' The Chancery Types of Italy and France ', *Fleuron*, III, 1924, pp. 23-52.

[4] Now chiefly remembered for his Greek *New Testament* which remained the ' Received Text ' for over three centuries. (At school I and my contemporaries studied the Greek Testament in a reprint subtitled " Textus Stephanicus, A.D. 1550 ". *C. B.*) In searching for an italic to accompany such faces as Bembo or Centaur, modern designs have usually reverted to the originals of Arrighi or Blado.

psychologist. Might not publishers benefit by similar preliminary investigations, instead of trusting to the conventional or impressionistic judgment of their production-manager or printer?

Mathematical Publications. The needs of a mathematical publication are somewhat specialized. The italic letters used as algebraic symbols seem to have become pretty well fixed during the 19th century when textbooks were for the most part printed in a ' modern ' face. Hence many of our readers thought the *g*, *v*, *w*, *x*, and *y* as used in this journal " looked like an infringement of the customary symbols ". The symbols suggested in *The Printing of Mathematics* are certainly superior; but even they are not wholly satisfying to the statistical psychologist. For example, the ' bold face series ' there proposed (which the statistician would often need in matrix work) do not belong to the same type face nor do they correspond with the ordinary letters, as may be seen by comparing the C's, Q's, v's, and w's. Moreover, a statistical journal has additional problems of its own. Modern statistics requires Greek letters for the ' population ' corresponding in slope and x-height to the roman letters used for the ' sample ', and a variety of symbols far wider than that of the available alphabets. For these and other reasons Professor Kendall holds that there can be " no general standardization of symbolism, at any rate for some time to come ". On the other hand, he agrees that there might be " a case for standardizing symbols in a particular branch " (26, p. 8). These, however, are questions that we hope to take up at a later stage in our inquiries.

A Note on the Psychology of History. In our paper (p. 44 above) we ventured to suggest that the history of typography might provide instructive illustrations of the broader psychological principles that underlie the course of history. Each of the great philosophers of the past has produced his particular ' philosophy of history ': each has seen in it some kind of pattern, usually conforming with his special brand of metaphysics. The professional historian has, needless to say, usually scouted such generalized interpretations. Elsewhere, however, one of the present writers has argued that what is required is not a philosophy but a *psychology of history*, based on a direct analysis of empirical data. The historical study of typography would, we believe, admirably exemplify, in simple, concrete, and readily accessible form, most of the general principles that have been invoked to explain the larger course of events: the ' swing of the pendulum ', the ' threefold dialectical movement ' beloved of Hegel and Marx, the ' evolutionary pattern ', the ' patternless pattern ' (" just one dam' thing after another ", as Professor H. A. L. Fisher maintained), and finally the economic, the technological, and the cultural patterns. The study of such trends is primarily a problem for the psychologist, a problem which, perhaps because of its complex nature, the psychologist himself has hitherto left severely alone. Nevertheless, it is our firm conviction that, by starting with the history of relatively simple forms of human activity, such as the various arts and crafts, to which typography forms an excellent introduction, an illuminating series of researches might well be undertaken, which would yield valuable data and instructive conclusions both for the historian and for the social psychologist.

REFERENCES

1. Erdmann, B. and Dodge, R. (1898). *Psychologische Untersuchungen über das Lesen auf experimenteller Grundlage.* Niemeyer.
2. British Association for the Advancement of Science (1913). *Report of Committee on the Influence of Schoolbooks upon Eyesight.* Murray.
3. Burt, C. (1921). *Mental and Scholastic Tests.* King.
4. Parsons, J. H. (1922). ' Report of committee appointed to select best faces of types for Government printing ' (Review). *Brit. J. Ophthalm.*, VI, 475-9.
5. Updike, D. B. (1922). *Printing Types: their History, Forms, and Use.* Harvard University Press.
6. Burtt, H. E. and Basch, C. (1923). ' Legibility of Bodoni, Baskerville, and Cheltenham type faces.' *J. Appl. Psychol.*, VII, 237-45.
7. Kerr, J. (1926). *The Fundamentals of School Health.* Allen & Unwin.
8. Beaujon, P. (1926). ' The Garamond types.' *Fleuron*, V, 131-82.
9. Tinker, M. A. (1928). ' The relative legibility of the letters, the digits, and certain mathematical signs.' *J. Gen. Psychol.*, I, 472-96.
10. Buckingham, B. R. (1931). ' New data on the typography of textbooks.' *Yearbook Nat. Soc. Stud. Educ.*, XXX, 93-125.
11. Paterson, D. G. and Tinker, M.A. (1929-32). ' Studies of typographical factors influencing speed of reading ': ' II. Size of type,' ' III. Length of line,' ' X. Style of type face.' *J. Appl. Psychol.*, XIII, 120-30, 205-19, and XVI, 605-13.
12. Thorp, J. (1931). ' Toward a nomenclature of letter forms.' *Monotype Recorder*, XXX, pp. 9-19.

13. Tinker, M. A. (1932). 'The influence of the form of type on the perception of words.' *J. Appl. Psychol.*, XVI, 167-74.
14. Tinker, M. A. (1932). 'Studies in scientific typography.' *Psychol. Bull.*, XXIX, 670-1.
15. Burt, C. (1933). *The Psychology of Art* (ap. *How the Mind Works*). Allen & Unwin.
16. Ovink, G. W. (1938). *Legibility, Atmosphere-value, and Forms of Printing Types.* Leyden: A. W. Sijthoff's Uitgeversmaatschappig.
17. Williams, E. D. *et al.* (1938). 'Tests of literary appreciation.' *Brit. J. Educ. Psychol.*, VIII, 265-84.
18. Burt, C. (1939). 'The factorial analysis of emotional traits.' *Character and Personality*, VII, 238, 254, 285-99.
19. Luckiesh, M. and Moss, F. K. (1940). 'Criteria of readability.' *J. Exp. Psychol.*, XXVII, 256-70.
20 Paterson, D. G. and Tinker, M. A. (1940). *How to Make Type Readable.* Harper.
21. Simon, O. (1945). *Introduction to Typography.* Faber & Faber.
22. Carmichael, L. and Dearborn, W. F. (1948). *Reading and Visual Fatigue.* Harrap.
23. Anon. (1950). 'Fifty years of typecutting.' *Monotype Recorder*, XXXIX, 1-28.
24. Morison, Stanley (1951). *First Principles of Typography.* Cambridge University Press.
25. Jennett, Seán (1951). *The Making of Books.* Faber & Faber.
26. Kendall, M. G. (1954). 'The projected dictionary of statistical terms.' *Bull. Internat. Statist. Inst.*, XXXIV, 3-15.
27. Chaundy, T. W., Barrett, P. R., and Batey, C. (1954). *The Printing of Mathematics.* Oxford University Press.

Note on Table of Type Faces. For the greater part of the nineteenth century printers had practically only one style of type at their disposal; today they can select a design from almost any period in the history of printing. The table opposite shows the most important of these faces, and is intended to illustrate the specimens chosen for our experiments and the criticisms made in the introspective comments.[1] To aid the reader in distinguishing the characteristics of the classes of type face indicated by the factorial analysis, we have included only those letters that vary widely from one fount to another; and, instead of following the usual alphabetical order, we have placed first those letters which have variants common to the largest number of families or groups. The nominal size is the same throughout, viz., 12 pt. In the text, where letters from different founts have been described or discussed, the exigencies of machine composition have forced us to set them in the type-face used for the rest of the article, namely, Times New Roman, *not* in the particular fount referred to.

[1] We regret that the printer has been unable to supply some of the type faces discussed in our article; the omissions relate chiefly to the less important founts, namely, Granjon, Didot, Centaur, and Gill Sans. Mr. H. W. Mortimer, a teacher who has been good enough to read the proofs, thinks that we should have included Gill Sans in our main experiments, "particularly when studying types for children's reading-books, for which it has achieved some popularity". However, as mentioned above (p. 32, footnote 2) our pilot inquiry indicated that, quite apart from their rarity as book faces, sans-serif designs by no means possess all the merits claimed for them. Here our conclusion is in keeping with that of H. R. Crosland and H. Johnson, who also found "serifed letters more legible than unserifed" (*J. Appl. Psychol.*, XII, 1928, p. 121). In the reading of *consecutive* print, as contrasted with the mere recognition of *isolated* letters, the serifs contribute appreciably towards the combination of the letters into a characteristic word-whole: as one type designer puts it, "they seem to aid the flow of the eye" (a phrase which is perhaps open to psychological criticism). In our experiments on numerals we certainly found Gill Sans better for *short* numbers (e.g., hours of trains, etc., as in railway time-tables, where they also facilitate economy of space); but they were less satisfactory for *long* numbers (as in mathematical tables), presumably because they do not combine so readily into distinctive groups. But we hope in the near future to publish a separate investigation on type for numerical publications, such as children's arithmetic books, tables in statistical papers and the like.

SPECIMENS OF TYPE FACES

IA1 Bembo

W M J Q R G e j g Th
v w y z k p m g h J Q W

IA1 Centaur

W M J Q R G e j g h y
v w y z k p m g h J Q W

IA1 Veronese

W M J Q R G N e j g z
v w y z k p m g h J Q W

IA2 Garamond

W M J Q R G N e j g u n
v w y z k p m g h J Q W A V R

IA2 Fournier

W M J Q R G Z A e b j g
v w y z k p m g h J Q W

IA2 Plantin

W M J Q R G C A P e j g
v w y z k p m g h J Q W

IA2 Ehrhardt

W M J Q R G C A E T U j g c b p q f r
v w y i j k p m g f J Q W R P B D

IB Caslon

W M J Q R G C A H U e
v w y z k p m g J Q W C Y A V H

IB Baskerville

W M J Q R G C A H e w g s
v w y z k p m g J Q W C Y K N T

IB Old Style

W M J Q R G C A H e w g s
v w y z k p m g J Q W C Y

IB Imprint

W M J Q R G C A e w g s
v w y z k p m g J Q W C Y

IB Times New Roman

W M J Q R G C e w g s c b p q
v w y z i j k p m h J Q W

IB Perpetua

W M J Q R G C A E T U j g c b p q f r
v w y i j k p m g f J Q W R P B D

IIA Bodoni

W M J Q R G C K e w j b t u n
v w y i j k p m t x g J Q W R

IIA Walbaum

W M J Q R K e w g j b t
v w y i j k p m t x J Q W R

IIB Bell

W M J Q R K e w j b t k
v w y i j k p m t x g J Q W

IIB Scotch

W M J Q R K e w j b t k
v w y i j k p m t x g J Q W R

IIB Modern 7

W M J Q R K U e w j b t h
v w y i j k p m t x g J Q W R

INTERNATIONAL CONFERENCE ON FACTOR ANALYSIS

The *Centre National de la Recherche Scientifique*, aided by a generous grant from the Rockefeller Foundation, is organizing a *Colloque International* on ' Factor Analysis and its Applications ', to be held in Paris from 11th to 16th July, 1955. The provisional programme is as follows:

11TH JULY. Inaugural addresses and report on the Uppsala symposium.
Professor L. L. Thurstone, ' Recent studies in factor analysis '.
Professor Sir Cyril Burt, ' Factor analysis: historical survey of methods and results '.
Professor H. Piéron, ' Le problème général de la recherche et de la nature des facteurs en psychophysiologie '.

12TH JULY. *Professor H. Hotelling*, ' Relations of the newer multivariate statistical methods to factor analysis '.
Professor H. Pineau, ' Remarques sur l'analyse factorielle de Hotelling et comparaison avec les méthodes centroïdes '.
Professor M. Yela, ' Résultats et signification psychologique de l'analyse factorielle '.
Mme. G. Bernyer, ' Les facteurs psychologiques: leur nombre, leur identification, leur nature '.
Professor M. Reuchlin, ' Facteurs obliques, facteurs orthogonaux, facteurs de second ordre, en psychologie '.

13TH JULY. *Dr. L. Guttman*, ' The Radex approach to factor analysis '.
M. J. M. Faverge, ' Analyse factorielle et combinaisons linéaires de variables '.
Dr. H. J. Eysenck, ' The validity of factor analysis '.
M. E. Schreider, ' Application de l'analyse factorielle à l'étude de la variabilité biologique '.
Dr. S. Ledermann, ' Application de l'analyse factorielle à l'étude de la mortalité '.

15TH JULY. *Professor G. Darmois*, ' Quelques résultats théoriques et leurs conséquences pour les applications '.
Professor P. Delaporte, ' Nouvelle méthode de statistique mathématique pour l'estimation des facteurs '.
Dr. A. H. El Koussy, ' Trends of research in space abilities '.
Professor T. Husen, ' Factor analysis of achievement tests '.
Professor E. A. Peel, ' The factor analysis of correlations between persons with independent determiners to identify factors'.

16TH JULY. *Sir Cyril Burt*, Summary of Conference. Final Discussion.

Vol. VIII Part I The British Journal of Statistical Psychology May 1955

BOOK REVIEW

Faster than Thought: A Symposium on Digital Computing Machines. Edited by B. V. BOWDEN London: Pitman & Sons. Pp. xx + 416.

MUCH of the calculation required in statistical psychology will in the near future be performed by electronic computers. Hitherto, psychologists who have devised working techniques for factor analysis and similar procedures have striven to keep them as simple as possible: usually they have had in mind a student or research worker equipped with only a hand machine or with nothing at all. This preference for easy methods—for graphical rotation instead of arithmetical rotation, simple summation instead of weighted summation, guessed communalities instead of correcting for specificity or applying maximum likelihood—has frequently provoked the criticisms of the professional mathematician, and is now gravely impeding the development of rigorous and efficient formulae. The time has therefore arrived for the psychologist to adapt his working methods to the improved facilities now available; and with that end in view he should familiarize himself with the nature and capabilities of these novel appliances. As an introduction to the subject the volume that Dr. Bowden has edited is at once the most up-to-date, the most readable, and the most authoritative.

Dr. Bowden begins by reminding us how, in every civilized country, the organization of industry and government is growing more and more complex. "The welfare state", he warns us, "can only be run efficiently on a diet of numbers; and as a result we seem to be fast becoming a nation of computing clerks." But the 'electronic brain' is "now waiting to come to the rescue". Not only can it calculate with astonishing speed; it can store, recall, and reorganize data far more accurately and promptly than a whole army of clerks. Unfortunately in Britain (so at least we are told) the social psychologists, the industrial psychologists, and those who advise on 'scientific management' have hitherto shown little sign of appreciating what can be accomplished by such means.[1] If Dr. Bowden's hopes are realized, this generation will soon be witnessing a second industrial revolution: in the first, the machine replaced men's muscles, so that every English workman now has on an average three horse-power to help him; in the next, it will largely replace his brain, and incidentally lighten the drudgery and boredom that now oppresses the white collar worker.

The book, which is profusely illustrated, falls into three main parts. The first deals with the history and theory of mechanized computation; the second with the commoner types of machine that have recently been built; and the third and longest with the more important fields of application.

The opening chapter, written in Dr. Bowden's liveliest style, presents a brief historical survey of calculating devices from the days when Sumerians stamped their accounts on clay tablets. To the psychologist the spasmodic way in which the whole project has developed is itself a problem of singular interest. The first effective adding machine was built by Blaise Pascal in 1642 at the age of 19, to assist his father, who was an officer of the French customs: it can still be seen in the *Conservatoire des Arts et Métiers*. Thirty years later, Leibniz exhibited before the Royal Society of London an invention of his own which could not only add but multiply. Nevertheless, two hundred years later the tables in the *Nautical Almanac* were still being calculated by officials at the Greenwich Observatory, and contained so many inaccuracies that few experienced navigators would use them.[2] However, shortly after the beginning of the nineteenth century, it occurred to a young Cambridge mathematician that fatal errors of this kind could be avoided if all tabular functions could be computed by machinery; and, although during his lifetime nothing practical emerged from his prolonged investigations, it is now quite obvious that he had successfully hammered out all the basic principles on which modern digital computers are built.

Charles Babbage was the son of a banker, and thus a man of means. Though he had left Cambridge without sitting for the Tripos, he became a Fellow of the Royal Society at 22, and at 36 was elected to the Lucasian Professorship of Mathematics, the Chair once held by Newton. This office he retained from 1829 to 1839 without giving a single lecture. Babbage was a highly original inventor. He is said to have devised the cow-catcher, the first speedometer, 'skeleton keys for unpickable locks', ingenious dodges for solving codes and ciphers, a valuable method of detecting

[1] Quite recently J. Lyons & Co. have installed 'Leo', an electronic office which works out the weekly wage-packets of their 7,000 London employees, analyses day by day the trend of orders in the various teashops, and is rented in its spare time to outside bodies. To cover developments in automatic control, a vice-president of the Ford Motor Company has coined the word 'automation' (cf. R. K. Geiser, *Conference on Automation and Industrial Development*, New York, May, 1954). In this country the Joint Committee on Human Relations in Industry, set up by the Department of Scientific and Industrial Research and the Medical Research Council has sponsored a small pilot research at Cambridge on automatic control in industry; and, while these pages are going through the press, P.E.P. has issued an instructive pamphlet on the subject ('Towards an Automatic Factory', *Planning*, XXI, 1955, pp. 64-84).

[2] We are told that Captain Smyth, R.N., while making a survey of the Mediterranean, fell in with a Spanish vessel, and, during the exchange of courtesies, presented the navigator with a handsomely bound copy of the *Almanac*. Captain Smyth who "sailed by rule-of-thumb", returned to Portsmouth quite safely: the Spaniard was never heard of again.

climatic cycles based on rings of tree-trunks, and finally the occulting lighthouse. His analysis of post office economics led to Rowland Hill's introduction of the penny post; and he even anticipated the experiments of Professor J. B. S. Haldane by spending ten minutes in an oven at a temperature of 265° Fahrenheit.

To cut the cogs for his computing machines with sufficient regularity, great improvements had to be made in lathes and precision-tools; and in workshop methods alone his achievements, as Dr. Bowden remarks, must have more than justified the grant which the Government gave him towards constructing his earlier models. To feed in the figures for his computing mechanisms to work with, he hit upon the idea of using the punched cards that had just been invented by a French weaver, Joseph Jacquard, for controlling his looms.[1] Unhappily, and not always through faults of his own, hardly any of the machines that Babbage designed and started were actually completed; dexterous as he was in handling tools and formulae, he had no flair for explaining his ideas to others and was wholly devoid of the tact and good temper needed to cope with government officials and scientific critics. He died, a disappointed visionary, in 1872; and his workshops were found littered with what a contemporary described as " the wreckage of a brilliant career, and bits of machine which will doubtless remain for ever a mere theoretic possibility ".

In his own eyes his greatest feat was the invention of a new algebra to describe the logical relations between the various parts of his machinery. The symbolism proposed was highly ingenious. But Babbage himself was too engrossed in applying his principles to publish any clear or comprehensive exposition of them. The best account both of his methods and of his machines we owe to another extraordinary personality, Ada Augusta, Countess of Lovelace.

Early in 1815, Lord Byron married " a young heiress and savante " named Miss Milbanke. In his love letters he calls her his " Princess of Parallelograms ". In *Don Juan* [2] she figures as " the winsome Donna Inez, Don Josè's learned wife ". " Her favourite science was the mathematical." A year later " Don Josè and his lady quarrelled ", and " were for ever parted ". Meanwhile a child had been born—

> Ada, sole daughter of my house and heart,
> A child of bitterness, though born in love.[3]

" My voice ", he says elsewhere, " shall with thy future visions blend, To aid thy mind's development." And indeed her mind apparently inherited both her father's literary talents and her mother's mathematical abilities. As a girl she studied under Augustus de Morgan, Professor of Mathematics at University College, London. One day his wife took her to see ' Mr. Babbage's engine '. " While the rest of the party " (says Mrs. De Morgan) " gazed at the instrument with the expression savages are said to wear on first seeing a clock or a gun, Miss Byron, young as she was, saw the great beauty of the invention and understood its working."[4] A few years later she married William King, afterwards first Earl of Lovelace. She and Babbage remained close friends; and of all his contemporaries she alone seems to have really grasped what he was trying to do.

In 1840, Babbage gave a series of lectures at Turin. They were attended by L. F. Menabrea (an officer of the Italian Military Engineers and in later years one of Garibaldi's generals) who wrote out a detailed ' Sketch of the Analytical Engine invented by C. Babbage '. Lady Lovelace translated his paper into English, supplementing the description with valuable mathematical notes and first-hand explanations, and drawing up an intricate chart or ' programme ' for computing the Bernouilli numbers by way of illustration. Later she and Babbage worked out an elaborate mathematical system of betting on horses, and endeavoured by its means to raise funds to build a still better machine. Both lost heavily. Shortly afterwards, at the early age of 37, she died, and was buried beside her father at Newstead Abbey. Dr. Bowden has had the happy idea of reprinting the whole of her memoir in an Appendix of sixty pages.[5]

[1] A few years after Babbage's death the same idea occurred to Dr. Hollerith, then in charge of the American Bureau of the Census. Although by law a U.S. census had to be taken every ten years, the work had grown so rapidly that with existing methods it would have taken twenty years to complete the requisite calculations. Hence the ' Hollerith sorter ' with which every psychologist is nowadays familiar.

[2] The story is told in the first thirty cantos, as an awful warning to any who would " wed a walking multiplication table ", with the final moral:

> Oh, lords of ladies intellectual,
> Inform us truly, have they not henpeck'd you all?

The footnotes to Murray's edition of 1837 explain the rather cryptic allusions. See also ' *Fare Thee Well*, March, 1816 ', the manuscript of which, says Murray, " is blotted all over with the marks of tears ".

[3] *Childe Harold's Pilgrimage*, III, i, cxviii. It is curious that no editor, commenting on the poet's repeated speculations about the " destiny of his only child ", ever tells the reader what her destiny actually was.

[4] See this *Journal*, IV, 1951, pp. 198 f. As Dr. Bowden remarks, " the attitude of the average intelligent man towards mathematical devices has changed very little during the last hundred years ".

[5] Babbage himself considered Lady Lovelace's account of his machine to be the best available, and noted that she had corrected an error in his own recurrent formula for computing the Bernouilli numbers.

A passion for mathematics, horse-racing, and foreign adventure appears to have run through her family for two or three generations. Her daughter, a brilliant mathematician and Arabic scholar, married Wilfred Scawen Blunt, the poet, and sought to introduce Arab blood into the race horse (see *The Authentic Arabian Horse*, 1942, by Baroness Wentworth, her granddaughter). When in a broadcast talk I mentioned Lady Lovelace and her mother and daughter as members of a remarkable mathematical family overlooked by Galton, I received a letter from another of her descendants telling me that several living members were equally gifted, and that one happened to be working for a well known manufacturer of computing machines.

The subsequent history of computing machines from the days of Babbage to those of the earliest electronic computers is dismissed rather briefly [1]; and the remainder of Part I consists in a comprehensive and lucid account of the internal organization of the modern high speed instrument and of the circuit components and other mechanisms on which its working depends.

Part II is devoted to the description of existing machines. Nearly a dozen different specialists contribute a chapter each on the latest types now working in this country and in the United States. The British psychologist in particular will be grateful for the full details about the construction and capabilities of the machines available at Birkbeck College, Imperial College, and the National Physical Laboratory.

The third and longest part of the book is concerned with the applications of such machines. It discusses at some length the services which they have rendered, or might render, in government offices, in business and commerce, in the fighting services, in crystallography, meteorology, dynamical astronomy, and in various branches of scientific engineering. To the psychologist the chapter of greatest interest will be that dealing with the solution of logical problems.

Once again the history of the basic idea goes back to Leibniz; in a well-known passage, he declared that, with the methods that he advocated, " wrangling would be as unnecessary between philosophers as it is between accountants: they would simply say: ' Let us calculate ' ". To take his own example, suppose ' man ' to be represented by the number 6; then the prime factors of this number may represent ' rational ' and ' animal ': thus the proposition ' man is a rational animal ' will be represented by analysing 6 into its factors, 2 and 3. And, with more complex subjects, an *ars combinatoria* will enable us to construct and compare all the relevant combinations of such factors.[2] In this way logic will become a kind of algebra.

What Liebniz only dreamt of, George Boole actually carried out. In 1848, Boole (professor of Mathematics at Queen's College, Cork) published his *Mathematical Analysis of Logic*, which was later expanded into his more celebrated *Laws of Thought*—a book that Bertrand Russell once hailed as the most momentous contribution to formal logic ever written. Its object was " to express the operations of reasoning in the language of a calculus ". From this it was a comparatively simple step to invent an apparatus that would apply the calculus. The first machine to be actually constructed was the ' logical piano ' designed by Stanley Jevons (another of De Morgan's students at University College and later Professor of Political Economy there): a woodcut of it forms the frontispiece to his book on *The Principles of Science*.[3] The instrument, as he tells us, was inspired by the work of De Morgan and Boole. Its operation assumes the logic of dichotomous classification (X and not-X, etc.); and the essential processes have an instructive analogy with the factorial analysis and factorial synthesis of qualitative data.

The statistical psychologist often finds it difficult to persuade his critics that the algebraic methods that he adopts are merely an application of rigorous logical principles: perhaps a more effective line of explanation would be achieved if he argued in the opposite direction, and tried to show, as Jevons does, that complex logical problems, starting with purely qualitative data, can be efficiently solved by the application of mathematical formulae. Let us therefore begin by comparing the procedure proposed by Jevons with that adopted by the statistical psychologist.[4] This will help us to a clearer understanding of the principles embodied in the more modern logical machines described by Dr. Bowden.

The logical reasoning which Jevons' machine is capable of performing is based on the principle that all purely logical relations may be expressed in terms of (1) conjunction and (2) negation: thus, ' *Either* A *or* B ' may be expressed by saying ' The conjunction of Not-A and Not-B does not occur '; ' *If* A, *then* B ' by saying ' the conjunction A and Not-B does not occur '; and so on (Jevons, *loc. cit.*, pp. 73 f.). And from this it would evidently follow that every statement, so far as it involves logical relations, can be expanded in *a logical sum of logical products* (the so-called ' Boolean expansion '). Such an expansion can be computed by methods akin to matrix multiplication, which consists essentially in forming product-sums. However, instead of the somewhat abstruse examples borrowed

[1] For this intervening period the reader may refer to David Baxandall's excellent article on ' Calculating Machines ' in the *Encyclopaedia Britannica* (14th edn., 1929). A machine which fully exploited the principles set forth by Babbage was not actually built until a hundred years after he had formulated them. This was the sequence-controlled calculator, known as Harvard Mark I, designed by Professor Aiken of Harvard in 1939 and put into service a few years later; its calculating elements consist of mechanical counters driven through electro-magnetic clutches which in turn are controlled by electromechanical relay circuits. The first all-electronic computer was the Electronic Numerical Integrator and Calculator (ENIAC) built at the School of Engineering, University of Pennsylvania, 1946: it contained 18,000 valves and 1,500 relays, and works several hundred times faster than the Harvard Mark I. Still more ingenious models have since been designed by Professor von Neumann and his co-workers at Philadelphia, and affectionately christened EDVAC, EDSAC, JOHNIAC, and MANIAC.

[2] Liebniz's proposals are contained in a rather rare tract, entitled ' Non inelegans Specimen Demonstrandi in Abstractis ' (Erdmann, *Leibnitii Opera*, 1840, I, pp. 94 f.). The most instructive passages are quoted by Jevons in the Preface to the second edition of his *Principles of Science* (1877).

The ' frame or engine ' with forty iron handles " for improving speculative knowledge by mechanical operations ", which the professor demonstrated to Gulliver at the ' grand academy of Lagado ', was presumably suggested to Swift by Leibniz's demonstrations to the Royal Society: Sir Walter Scott, however, supposed it to be based on Ramón Lull's *Ars Magna*.

[3] W. S. Jevons, *Phil. Trans.*, CLX, 1870, pp. 497 f.; *id.*, *The Principles of Science*, 1874, pp. 108 f.

[4] Cf. ' The Factorial Analysis of Qualitative Data ', this *Journal*, III, 1950, pp. 166-185.

by Jevons from physics or chemistry, let us start with a concrete problem so simple that its solution is obvious almost at once.[1]

Just before the war, a local education authority gave the following instructions to a recently appointed psychologist. " Pupils who are innately deficient in general intelligence or are backward in the essential subjects of the elementary curriculum by three or more years are regarded as educationally subnormal. All other pupils will be regarded as normal and will attend the ordinary elementary school. Pupils who are subnormal must be recommended for transference to the special school. The term ' pupil ' here includes any boy or girl between the ages of 5 and 14 who is physically fit for classroom instruction." The psychologist thereupon asked himself: " how precisely do I classify the children referred to me, and which of these classes am I to recommend as fit for a special school or for an elementary school respectively? "

Confronted with a problem such as this, we start, as Leibniz and Jevons would have us do, by expressing the relevant qualifications in terms of four components or ' factors ': (i) G, a positive or ' general ' factor, denoting the whole universe or *genus* of ' pupils '; (ii) X, a first ' bipolar ', denoting suitability for a ' special school ' or the reverse; (iii) A, another bipolar, specifying presence or absence of innate ' deficiency '; (iv) B, a third bipolar, specifying presence or absence of ' educational backwardness '. Thus, in theory, the entire *genus* will consist of $2^3 = 8$ conceivable classes of persons which we can number from i to viii.

We now require symbols to indicate the positive or negative aspects of these components. With literal symbols, we can use (with Jevons) X and x, or (with Woodger and others) X and $-X$, or (with Welton) X and $\overline{X}$. With numerical symbols we can enter in the row or column reserved for the component in question either (*a*) + 1 (for the positive term) and − 1 (for the negative term) or again (*b*) 1 (for presence) and 0 (for absence). It will be seen that these numerical alternatives correspond with the factorist's choice between a bipolar factor and a group factor. In the electronic machine the alternatives may similarly be represented in two different ways. (*a*) They may be indicated by a pair of contrasting signals, the mere absence of a signal conveying no information: this is the principle adopted in the old Morse code machines and in the machine built by the Telecommunications Research Establishment at Malvern. Or (*b*) they can be represented by a relay or a switch which is either ' on ' or ' off ' : this is the device now adopted by most electronic machines.

In my previous paper I used the second type of notation (0 or 1). Here I shall attempt to show how the first can be used. Adopting this convention we can specify the 8 classes by a ' bipolar ' matrix, such as the following:

Classes of Persons	i	ii	iii	iv	v	vi	vii	viii
Factor G	+1	+1	+1	+1	+1	+1	+1	+1
„ X	+1	+1	+1	+1	−1	−1	−1	−1
„ A	+1	+1	−1	−1	+1	+1	−1	−1
„ B	+1	−1	+1	−1	+1	−1	+1	−1

With the alternative notation we should obtain a matrix in terms, not of ' bipolar factors ', but of ' group factors '. But the bipolar matrix can evidently be rotated to the group factor form by means of the following pre-multiplier:

		Bipolar Factors			
		G	X	A	B
Group Factors	G′	1	0	0	0
	X′	½	½	0	0
	A′	½	0	½	0
	B′	½	0	0	½

Out of the eight conceivable classes deducible from these three ' bipolar factors ' the regulations regarding special schools exclude a certain number by imposing the requirement $X = A \vee B$ (i.e., that X must be either A or B). The next step, therefore, is to determine what classes are covered by the requirement ' either A or B '. The requisite operation can be carried out by first rotating the bipolar matrix as above described, and then applying the summation premultiplier $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$ to obtain the ' logical sum ' of the two lower rows resulting from the rotation, namely A′ and B′. The type of summation required will follow the usual rules, except [2] that $(+1)+(+1) = +1$: (the procedure is in some ways analogous to ' unweighted summation ' in factor analysis). The premultiplier $\begin{bmatrix} -1 & 2 & 0 \\ -1 & 0 & 2 \end{bmatrix}$ will then convert the three one-or-zero rows back to bipolar form, thus yielding two bipolar rows.

[1] It should be noted that ' or ' is taken in a non-exclusive sense: as Mill puts it, " when we say X is ' either a knave or a fool ' we by no means imply that he cannot be both ". Thus our ' group factors ' may overlap. Dr. Bowden's example therefore seems unfortunate, since it tacitly rules out the *joint* possibility, whereas his machine obviously does not.

[2] The apparent paradox has a perfectly logical explanation: +1 in the new row A and cols. i, ii, and iii means " all in classses i, ii, and iii are A "; +1 in the new row B means " all in subclass i are B ". The logical sum states ' all in classes i, ii, and iiii are either A or B ' (in the non-exclusive sense). And in each case we must again use ' +1 ' to denote ' all '.

For the last step we note that the relation of equivalence is expressed by the operator $[\frac{1}{2}, \frac{1}{2}]$. On using it to pre-multiply the two bipolar rows just obtained, we get a single row, whose values contain the solution to our problem, namely:

Subclass	i	ii	iii	iv	v	vi	vii	viii
Final result	1	1	1	0	0	0	0	−1

These results indicate that, with the requirements imposed, only the first three classes, $GX(AB)$, $GX(A\bar{B})$, $GX(\bar{A}B)$, and the last, $G\bar{X}(\bar{A}\bar{B})$, satisfy the regulations: the former, as we see, specify those persons that are fit for a special school (X), the latter, those that are not ($\bar{X}$). The four remaining classes (those marked zero) are forbidden by the regulations, either explicitly or implicitly.

In practice most readers would doubtless solve an elementary problem like this almost intuitively, without recourse to symbols and conscious inference.[1] But with the blind working of an artificial apparatus a step-by-step routine is unavoidable. To demonstrate the possibility of such an instrument, Ferranti, Ltd., have built an electronic contrivance capable of dealing with three components: at Harvard a larger machine will deal with as many as twelve.[2]

With the method I have described, it will be seen, the last step of all involves scanning the final results for each possible combination. Three components yield only 8 combinations; but 12 would yield as many as 4,096; and twenty over a million. Thus the number to be scanned rises swiftly to an amount that is quite impracticable even for the most up-to-date high-speed appliances. Fortunately, when the problem is of great complexity, it is often sufficient to obtain just one positive solution, without discovering all. In that case an alternative principle may be exploited—that of feed-back. Constructed in this fashion the machine can work 50,000 times faster; and its behaviour is virtually that of what is known as a discrete Markoff process.[3] Once again Ferranti, Ltd. have designed a machine based on this principle, which is also capable of furnishing *every* solution, when the problem is of a comparatively simple type.

Most of the logical problems handled by such appliances are, like the example I have used, problems to be solved by deduction: the data of the problem offer us the relevant classificatory components or 'factors' ready to hand. But in the natural sciences such classifications have to be discovered first of all. In psychology, this preliminary inductive stage constitutes one of the special tasks of factorial research—that of factor analysis as distinct from factor synthesis. It is to be hoped, therefore, that in the near future both designers and programmers will devote attention to these more exacting demands.[4]

The machines that seem chiefly to have hit the popular imagination are those that play games. As long ago as 1769, Baron Kempelen exhibited a chess-playing automaton which was able to defeat the Emperor Napoleon. It was described in some detail by Edgar Allân Poe; but, according to the story, the hoax was unmasked when somebody shouted 'Fire', and the legless dwarf concealed inside struggled frantically to escape. Babbage designed an apparatus which could play 'noughts and crosses'. And many readers may have tried to beat the machine called 'Nimrod' at the game of Nim, when it was installed at the Science Museum during the Festival of Britain. Dr. Bowden describes computers that will play both chess and draughts, the former rather badly, the latter quite well.

Encouraged by these remarkable feats of scientific engineering, physiologists, psychiatrists, psychologists, and philosophers (chiefly psychologists who are behaviourists and philosophers who are logical positivists of the physicalistic persuasion) have recently proclaimed that we are now wholly justified in regarding the human mind as no more than a highly developed physical machine. Henceforth we may confidently dispense with what Professor Ryle has dubbed 'the ghost within the machine'. Allowance must of course be made for differences in complexity: while the most elaborate computer contains no more than 5,000 valves, the average brain includes over 10,000,000,000 nerve cells (about six times the population of the globe). That being so, it is argued, we may safely attribute the obvious disparities between the two to the inequality in initial equipment. In short, "mind and mechanism differ only in degree and not in kind"—a conclusion which raises the perplexing questions discussed by Dr. Bowden in the closing chapter of his book.

He begins by comparing the actual performances of the most efficient machines with those of the most efficient human computers, and presents a highly instructive survey of the accomplishments

[1] The reader who finds the foregoing excessively simple may try one of those cited by Jevons, which contains the same number of components: "According to the prospectus members of the public could subscribe to the company by purchasing bonds, shares, or both. The directors are all bond-holders or share-holders but not both; and all the bond-holders are directors. What conclusion can be drawn?" We are told that, of 150 students, only 6 gave the answer. It may be noted that Jevons' instrument is capable of inductive as well as deductive inference.

[2] Cf. W. Mays and D. G. Prinz, 'A Relay Machine for the Demonstration of Symbol Logic', *Nature*, CLXV, 1950, Feb. 4.

[3] See this *Journal*, IV, 1951, p. 194.

[4] Using the method of principal components, Dr. Wrigley and Dr. Neuhaus have factorized a 7 × 7 correlation matrix in about 10 minutes' machine time ('A Refactorization of the Burt-Pearson Matrix with the Ordvac Computer', this *Journal*, V, 1952, pp. 105-108). But no one as yet has attempted solutions that depend on 'correction for specificity', 'simple summation', 'group factor analysis', or the use of 'subdivided factors'. To illustrate the unexpected difficulties of statistical work, Dr. Bowden prints a chart for 'part of a multiple repression (*sic*) analysis', but does not enter into detailed explanations.

and methods of Professor A. C. Aitken, F.R.S., and Mr. William Klein, of the Mathematisch Centrum, Amsterdam. On the whole, it would seem, the machine has the advantage as regards complexity of problem, memory for figures, and speed of solution. These happen to be the features in which 'calculating prodigies' themselves outdistance the ordinary man. On the other hand, the human expert possesses at least two advantages over the machine: first, he can choose or adapt his method according to the accidental peculiarities of each problem, and thus improvise short cuts in a way that is quite beyond the scope of any machine; secondly, he is endowed with a power of 'schematic apprehension', which no one has yet succeeded in imparting to an instrument: whereas the human calculator can grasp and manipulate whole patterns or *Gestalten*, the machine is compelled to work explicitly step by step.[1] *Prima facie* this suggests a difference of quality rather than of mere complexity or degree.

To examine the pros and cons of this ancient controversy would take us too far afield.[2] But it is worth while hearing the verdict of the engineers. As Dr. Bowden's quotations show, most of them are inclined to deprecate "the attempt to infer complete similarity from partial similarity, however striking". He himself points out that, although the engineers can produce a mechanized Barlow (who computed tables of squares and powers), and a mechanized Briggs (who compiled tables of logarithms), they have not even hinted at the possible production of a mechanized Napier (who *invented* logarithms). And at least one impartial critic, speaking as a philosopher as well as a psychologist, emphatically agrees. "There was once", says Professor Mace, "a monk who invented a machine which could prove the existence of God. The monk, however, was cleverer than the machine: for, up to the present, no machine has invented a monk who could prove anything at all."[3] Perhaps we may allow Lady Lovelace to have the last word (the italics are her own). "The analytical engine has no pretensions to *originate* anything; it can *merely do what we order it to perform.*"

CYRIL BURT.

[1] Professor Aitken, in the course of a broadcast in which Dr. Bowden and I were allowed to question him, remarked that the process seems to imply "a compound faculty, which, so far as I am aware, has never been exactly described". He suggests that "the analogy of the musician might help: a violinist concentrates, when he is playing, on the tune as a whole not on the separate notes or the fingering—except when he gets momentarily into a tangle".

[2] I have discussed the question in greater detail at the close of a recent paper in *Brit. J. Educ. Psychol.*, XXV, 1955, pp. 18 f.

[3] Preface to W. Sluckin's *Minds and Machines* (Penguin Books, 1954), p. 9. The reference is, I take it, to the earliest recorded logical machine, the *Ars Magna* (*c.* 1300) of Ramón Lull (a Spanish missionary and mystic who hoped that his proofs would convert the Moslems, but he was stoned to death by them at the age of 80). Francis Bacon, however, pronounced that it was a *methodus imposturae*.

Vol. VIII
Part II

The British Journal of Statistical Psychology

November
1955

THE DETERMINACY OF FACTOR SCORE MATRICES WITH IMPLICATIONS FOR FIVE OTHER BASIC PROBLEMS OF COMMON-FACTOR THEORY[1]

By LOUIS GUTTMAN
Israel Institute of Applied Social Research

I. *Introduction.* II. *The Problem of Determinacy.* III. *Supplementary Problems*

I. INTRODUCTION

1. *Problem.* Common-factor analysis—in the sense of Spearman, Thurstone and others, begins by considering the scores of a given population of N individuals on each of n quantitative variables from a given universe of content. Its goal is to express the observed scores as linear combinations of scores on common- and unique-factors.

It is typical of factor analysis procedures that they approach their goal only indirectly. If ξ is the observed score matrix, a *direct* analysis would consist in finding factor score matrices η and ζ such that

$$\xi = A\eta + \zeta \tag{1}$$

where A is some real matrix of common-factor loadings. Instead of a direct resolution into factor scores such as (1), indirect procedures are used, which pivot on the observed correlation matrix, say R. Real matrices A, L, and U are sought, where L is Gramian and U is diagonal, such that

$$R = ALA' + U^2. \tag{2}$$

As is well known, unless further restrictions are imposed, infinitely many solutions always exist for the right member of (2). There are different schools of thought as to the conditions to be imposed on U, especially on the rank of $R - U^2$; this is the well-known problem of 'communalities'. There are also different schools of thought as to what conditions should be imposed on A and L; this is the problem of 'rotation of axes'. Irrespective of their differences, each school in practice seems to regard its main job as completed for a given ξ when A, L, and U have been computed for (2) to satisfy that school's particular criteria for communalities and rotations.

The purpose of the present paper is to explore the extent to which the indirect analysis of the scores in ξ via (2) is equivalent to a direct analysis of the type (1). Part of this problem has been studied previously—from a somewhat different point of view—by other writers, notably E. B. Wilson, Godfrey Thomson, and Walter Ledermann [see discussion and references in 15, pp. 371 f.]. In the present paper, we shall establish the existence of solutions to (1) given (2), and a set of necessary and sufficient conditions for the construction of all possible solutions. This will enable us to go further and examine in detail some basic questions concerning the scientific meaning of common-factor analysis.

2. *Sufficiency Versus Necessity: The Six Problems.* The reason why (2) has entered the picture of factor analysis is that it is a *necessary* condition for (1). If A, η, and ζ are given such that (1) is satisfied, if L and U^2 are the covariance matrices of the common- and unique-factor scores respectively, and if the unique-factor scores have zero correlations with each other and with all the common-factors, *then* (2) necessarily follows.

[1] This research was facilitated in part by an uncommitted grant-in-aid from the Behavioral Sciences Division of the Ford Foundation, and in part by a grant for methodological work from the Lucius N. Littauer Foundation.

Necessity is not the same thing as sufficiency. Computations concerned only with (2) yield only matrices A, L, and U^2. To reach the goal of the factor analysis requires one to ask further: does there exist a pair of score matrices, η and ζ, to satisfy (1) with the given A and with the covariances specified by the given L and U^2 ? If yes, how many such pairs of score matrices exist for the fixed A, L, and U^2? That is, to what extent is (2) *sufficient* for establishing a solution to (1)?

This existence problem turns out always to have an affirmative answer, but also to have *too many* affirmative answers, especially when $U^2 \neq 0$. Furthermore, for finite n, these alternative answers can differ widely among themselves. We shall establish conditions under which (when $N = \infty$) there will be essentially but one answer in the limit as $n \to \infty$.

The multiplicity of solutions for η and ζ when n is finite—even when A, L, and U^2 are fixed—has important implications for the psychological meaning of common-factor analysis, as well as for the computing procedures conventionally used. If we denote as ' Problem I ' the determinacy of η and ζ, solving Problem I throws light on the five further basic problems listed below. These are discussed more fully in the last part of the paper.

Problem II. Communalities. The problem of the ' unknown communalities '—or equivalently, of the choice of U^2—turns out to be related to Problem I. We show a unique solution is definable under certain conditions as $n \to \infty$.

Problem III. Meaning of factors. It has been thought by many that the meaning of common-factor scores in η could be ascertained from A in (1), or factor variables in η could be *named* according to the observed variables in ξ which have high loadings on them. This may be illusory in the light of the answer to Problem I, for η itself can have widely different simultaneous solutions no matter what the structure of a fixed A may be.

Problem IV. ' Inverted ' factor analysis. Merely considering the case where $N = \infty$, as for example when the variables in ξ are continuous, raises questions, which may not have been previously apparent, in respect of the proposal to ' factor ' people rather than variables.

Problem V. ' Second-order ' common factors. When L in (2) is not a diagonal matrix, some researchers have proposed ' factoring ' it in turn like R. Solving Problem I shows that ' second-order ' scores may be even more indeterminate than η and ζ, especially when the number of common-factors is small, so their use in the quest for parsimony may be quite illusory.

Problem VI. Rotation of axes. If η by itself is indeterminate, any rotation into an $\bar{\eta}$ remains just as indeterminate, no matter what $\bar{A}$ may result from A. Seeking meaningful factor scores merely by a correlational analysis, as implied by rotations, may be fruitless; for this further observations or experiments beyond ξ and R may be essential. Since many fundamental and useful properties of a common-factor space can be shown to exist and can be studied without specifying a particular set of reference axes [8, 9, 11], the rotation problem may have diverted attention from what can really be learned by a correlational analysis alone.

We shall actually solve Problem I for the case where the covariance matrix U^2 is not necessarily diagonal, or where the so-called β-law of deviation holds [9, p. 308]. The restriction of U^2 to a diagonal matrix turns out not to affect the existence and determinacy of possible solutions. Of our six problems, only Problem II is devoted exclusively to the special case of a diagonal U^2.

The mathematics of factor analysis has usually been cast in terms of matrix algebra. More generally appropriate is the linear algebra of an abstract Euclidean vector space which we shall adopt in this paper.

II. THE PROBLEM OF DETERMINACY

3. *Notation for Score Matrices.* Let x_j denote the jth variable selected from the universe of content, and x_{ji} the observed score of individual i on x_j. It has been customary to regard the x_{ji} as elements of a finite real matrix of order $n \times N$. However, since this implies that the observed variables are discrete for the given population, we shall prefer instead to use a notation that will allow for both the discrete and the continuous cases, for finite and for infinite N (countable or not countable). To this end, we shall define the *score matrix* ξ and its transpose ξ' to be respectively

$$\xi = \begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_n \end{bmatrix}, \quad \xi' = [x_1' \, x_2' \ldots x_n']. \tag{3}$$

That is, ξ is a single column with *statistical variables* as elements, and ξ' is the same elements written as a row. We define the *inner-product* of two elements x_α and x_j to be their covariance, and denote this by $x_\alpha x_j'$. While neither x_α nor x_j is a real number, their product $x_\alpha x_j'$ is. In particular, the variance of x_j is $x_j x_j'$.

With this convention as to inner-products, we can form the product $\xi\xi'$ by the usual rule of matrix multiplication. Clearly, this is a real covariance matrix of order n; we shall denote it by R:

$$R = \xi\xi' = [x_\alpha x_\alpha']. \tag{4}$$

It is customary to assume all observed variances to be equal to 1:

$$x_j x_j' = 1 \qquad (j = 1, 2, \ldots, n). \tag{5 a}$$

In this case, R in (4) is the *correlation* matrix for the variables in ξ. Convention (5 a) is not essential to this paper; but we desire to restrict ourselves to the case where no diagonal element of R vanishes or all observed variances are positive:

$$x_j x_j' > 0 \qquad (j = 1, 2, \ldots, n). \tag{5 b}$$

Since (5 a) can always be attained when (5 b) holds (by dividing through by standard deviations), the reader is at liberty to interpret R in this paper either as a correlation matrix or merely as a covariance matrix, according as he wishes to adopt (5 a) or only (5 b).

One consequence—though not the most important—of our inner-product convention is that the origins of measurement of the x_j need not be specified ; the means are arbitrary. There is, however, no loss of generality in assuming the means to be zero for the problem of the present paper. When N is finite, a score matrix such as ξ in (3) can be regarded as a partitioned real matrix, each element representing a row of N real numbers. Our rule for post-multiplication of score matrices is the same as the usual one for partitioned real matrices, except that the usual result is divided through by N. Since we also wish to deal with variances and covariances for $N = \infty$, the usefulness of our modified post-multiplication rule will be apparent; it allows for general *integrals* (Stieltjes-Lebesgue) of products as well as for ordinary product-sums. This is the basic motivation for our inner-product convention. Here we shall always consider n finite. Although we shall also be interested in certain limits as $n \to \infty$, any given value of n is finite.

4. *The Vector Space Concept.* When dealing with score matrices, in view of our definition of an inner-product, not all rules for real matrices (i.e., matrices with real numbers as elements) need apply.

For example, we do not define the product $\xi'\xi$; but we shall also have no need for such a definition. On the other hand, if M is any real matrix of order $m \times n$, then the product $M\xi$ is defined by the ordinary law for a matrix product to yield a new column score matrix of m elements; linear combinations of variables yield variables. Similarly, if α and β are two column score matrices of n elements each, then the sum $\alpha+\beta$ is formed by the ordinary law for addition of matrices to yield a new column score matrix of n elements, for the sum of two variables is a variable. In short, column score matrices are to be pre-multiplied only by real matrices and post-multiplied only by row score matrices (in the latter case, the product being a real matrix of covariances). The remaining rules of matrix manipulations will hold for the needs of this paper.

In effect, we are regarding our statistical variables as points or elements of a *function space* (cf. 13, chap. V]. Such a space is a special case of a *Euclidean vector space* [1, p. 189]. The *norm* (length) of each of our elements is taken as its standard deviation, and the *inner-product* of any two of our elements is their covariance.

The algebra we need turns out to be completely appropriate to any finite dimensional (n finite) Euclidean vector space. It is actually more convenient to develop the algebra as for the general abstract case, and then *interpret* it for our statistical problem. Lemmas and theorems will usually be stated here then as for an n-dimensional Euclidean vector space. When we wish to analyse the general case, we shall speak, for example, of ξ as a *column of n vectors*. For our statistical case, ξ is a score matrix, or column of n statistical variables. The adaptability of the notation of matrix algebra to score matrices carries over—to the same extent—to columns of vectors from any Euclidean vector space.

One departure we shall make from a convention usually adopted for Euclidean vector spaces concerns the concept of a *zero vector*. By the zero vector, denoted by 0, is meant that vector such that $x+0 = x$ for all vectors x in the space. The convention is usually made that all vectors with *zero norms* shall be set equal to 0. We shall prefer to denote a vector with a zero norm by 0*, and allow it to be possibly different from 0. In our statistical case, a zero norm means a zero standard

deviation, or a variable 0* is *constant almost everywhere* (i.e., for almost all individuals), or equals *zero almost everywhere* if the mean is zero. It may be important in some cases to distinguish between a variable which is identically zero and one that merely has a zero variance, and we shall leave room for this distinction.

A vector with a zero norm will be called here a *null* vector. Clearly the vector 0 is a special case of a null vector. What all null vectors 0* have in common with the vector 0 is, that if x is any vector, then the inner-product of x with 0* vanishes: $0^*x' = 0$. This follows from Schwarz's inequality [1, p. 190]. A *column* of null vectors will be denoted below by θ.

A distinction between a Euclidean space and a Euclidean vector space is that in the former there is but one null vector, namely 0. As is customary, we shall use the symbol 0 to denote several different things: the real number zero, the zero vector, a zero matrix of real numbers, and a column of zero vectors. The context should always make clear which of these is implied.

5. *The Basic Hypotheses of Factor Analysis.* Given an observed score matrix ξ, part of the basic hypothesis of the Spearman-Thurstone factor theory is that hypothetical score matrices η and ζ exist, as well as a real weight matrix A, such that: each y_k of η has a correlation of zero with each z_j of ζ, or

$$\eta\zeta' = 0, \tag{6}$$

and (1) is satisfied.

If A is of order $n \times q$, then η is composed of q variables y_k which are called *common-factors*. ζ necessarily has n elements z_j in (1), although some of these may have zero variances; these n variables we shall call here *deviant-factors*. And, following a terminology suggested in [9, p. 308], we shall say that any η and ζ satisfying (6) obey the *β-law of deviation*.

The covariance matrices of the q common-factors and the n deviant-factors will be denoted by L and U^2 respectively:

$$L = \eta\eta', \qquad U^2 = \zeta\zeta'. \tag{7}$$

The Spearman-Thurstone theory requires further that ζ follow the *γ-law of deviation*, in the terminology of [9]. This means that $z_g z_j' = 0$ whenever $g \neq j$, or U^2 in (7) is a diagonal matrix. In such a case, if we denote the jth main diagonal element of U^2 by u_j^2, we can use the conventional notation for a diagonal matrix to write:

$$U^2 = [u_1^2, u_2^2, \ldots, u_n^2]. \tag{8}$$

When both the β- and γ-laws are satisfied, η and ζ are said to obey the *δ-law of deviation* [9, p. 308]. Following Thurstone, we shall then call z_j and u_j^2 the *unique-factor* and *uniqueness* respectively of x_j.

The solution to our sufficiency Problem I turns out to be essentially the same for the general β-law as for the more specialized δ-law. So we shall not restrict U^2 necessarily to be a diagonal matrix.

If L is a diagonal matrix, then the y_k are said to be orthogonal to each other; otherwise they are called oblique. We shall treat the general case of any L, diagonal or not.

It has been customary to assume the variances of the y_k all to be equal, and hence to set them all equal to 1. In some recent developments [9, 11], it appears desirable to express the y_k with unequal variances. So we shall treat the general case of arbitrary variances, and not necessarily assume the main diagonal elements of L all to be equal, nor in particular equal to 1.

The case of (1) usually studied is where A is of rank q and the elements of η are linearly independent, or L is non-singular. Since no extra work turns out to be involved thereby, we shall place no restrictions in general on the ranks of A and L. Of course, if the rank of A is r, then from the order of A it must be that $q \geqslant r$ and $n \geqslant r$.

Post-multiplying both members of (1) through in turn by η' and ζ' and using (6) and (7) show that

$$\xi\eta' = AL, \qquad \xi\zeta' = U^2 \tag{9}$$

Formulae (9) are well known for the δ-law; they hold more generally for the β-law.

The basic necessary condition (2) follows from post-multiplying both members of (1) by ξ' and using (4) and (9). Condition (2) involves no score matrix of (1) directly. It is well known for the case of a diagonal U^2, but holds again for the more general case of the β-law.

6. *The Existence of Solutions.* Hypotheses (1) and (6) by themselves do not serve to pin down q, r, or U^2, and certainly not A and L. Even when U^2 is restricted to being a diagonal matrix, this does not improve the situation. While their procedures may assume

many forms, in effect most factor analysts begin by assuming the δ-law, and seek a diagonal U^2 that will leave $R-U^2$ Gramian.

Suppose some decision on U^2 has been reached, and that for this U^2, the rank of $R-U^2$ is r. Factor analysts then proceed to 'factor' $R-U^2$ by any of several methods, to obtain a matrix A of order $n\times r$ and of rank r (that is, $q=r$), and a non-singular Gramian matrix L of order r, such that (9) is satisfied. The most general formulae for computing such an A and L from $R-U^2$ are given in [6, p. 12] and again in [7, p. 210]. These formulae hold whether U^2 is diagonal or not.

Our problem is: to what extent is satisfying (2) sufficient for establishing a solution to (1) that is also consonant with (6) and (7)? A general answer follows immediately from the following lemma.

Lemma 1. *Let ξ be a column of n vectors of rank p, each vector having a positive norm. Let $R=\xi\xi'$. If B and G are any real matrices such that*

$$R=\xi\xi'=BGB' \tag{10}$$

where G is a Gramian of order g and B is of order $n\times g$, then there exists a column ϕ of g vectors such that

$$\xi=B\phi, \qquad \phi\phi'=G. \tag{11}$$

For the proof, we first notice that the rank of G in (10) cannot be less than that of R (the rank of a product cannot exceed that of any factor in the product), and hence must be positive. Since G is Gramian, for any finite integer f not less than the rank of G there exists a real matrix F of order $g\times f$ such that

$$FF'=G. \tag{12}$$

From (10) and (12), $R=(BF)(BF)'$, so that the rank of R is that of BF, or BF must be of rank p. Since f is the number of columns in BF, clearly $f\geqslant p$. Then ξ can be regarded as imbedded in a Euclidian vector space of f dimensions. Let ω_f be the column of f vectors that form an orthonormal basis for this f-space; such a basis always exists (cf. 1, p. 193). Then there is some real matrix B_f such that $\xi=B_f\omega_f$. But since $\omega_f\omega_f'=I_f$ (the identity matrix of order f), the last equality implies $R=B_fB_f'$. BF and B_f both being 'factor' matrices of R and of the same order, there exists a real orthogonal matrix W (of order f) such that $B_f=BFW$ [4, p. 71]. Therefore, if we let $\phi=FW\omega_f$, then $B\phi=B_f\omega_f=\xi$, or the first part of (11) is satisfied. Also $\phi\phi'=(FW\omega_f)(FW\omega_f)'=FF'=G$, or the second part of (11) is satisfied. Hence, a ϕ always exists for (11) and Lemma 1 is established.

As an immediate consequence of Lemma 1 we have:

Theorem 1. *If A, L, and U^2 are given such that $R=\xi\xi'=ALA'+U^2$, where L and U^2 are Gramian and no diagonal element of R vanishes, then there exist two columns of vectors, η and ζ, such that $\xi=A\eta+\zeta$ and $\eta\eta'=L'$, $\zeta\zeta'=U^2$, and $\eta\zeta'=0$: that is, a solution to Problem I always exists.*

The proof consists of defining B and G in Lemma 1 by:

$$B=\|\,A\;I_n\,\|, \quad G=\left\|\begin{matrix} L & O \\ O & U^2 \end{matrix}\right\|, \tag{13}$$

where I_n is the identity matrix of order n, By direct multiplication, it is seen that $BGB'=ALA'+U^2$. Therefore, from Lemma 1 there exists a ϕ to satisfy (11). and (13). If L is of order q and U^2 of order n, then $g=q+n$. Let η be the first q elements of ϕ, and ζ the last n elements. Then

$$\phi=\left\|\begin{matrix}\eta\\ \zeta\end{matrix}\right\|, \quad \phi\phi'=\left\|\begin{matrix}\eta\eta' & \eta\zeta' \\ \zeta\eta' & \zeta\zeta'\end{matrix}\right\|. \tag{14}$$

But $\phi\phi'=G$ from (11), so comparing the second part of (13) with that of (14) shows that η and ζ satisfy (6) and (7). Furthermore, by direct multiplication according to the first parts of (13) and (14), $B\phi=A\eta+\zeta$ or—using the first part of (11)—we have verified that (1) is satisfied. This establishes the existence of η and ζ for Theorem 1.

Theorem 1 generalizes Theorem C of [7] to infinite N and to the general Euclidean vector space, and to the case where ζ is not regarded as fixed in advance.

It is of interest to inquire further *how many* solutions there are and how they are interrelated. From the proof of Lemma 1, it is obvious that ϕ cannot be uniquely determined if the rank of G exceeds the rank of R. Correspondingly, in Theorem 1, if the number of linearly independent elements in a pair of η and ζ exceeds p (the rank of R), more than one pair of solutions exist.

Results of considerable importance are obtained by looking closely into the cases where p is less than the sum of the ranks of L and U^2. This we shall do by *constructing* all possible solutions to Problem I. An incidental consequence will be an alternative proof of the statements in the preceding paragraph. More important, detailed information will be made available as to the relationship among alternative solutions and the implication these have for the foundations of common-factor theory. The Spearman-Thurstone theory is an important example of where p is less than the sum of the ranks of L and U^2.

7. *A General Construction Lemma.* Lemma 1 asserts the *existence* of a solution ϕ for (11), given (10). We should now like to show how all such solutions can be *constructed.* This is done in Lemma 2.

Lemma 2. *A necessary and sufficient condition that* ϕ *satisfy* (11) *in Lemma* 1 *is that it be of the form*

$$\phi = K\xi + C\omega + \theta, \tag{15}$$

where K *is a real matrix of order* $g \times n$ *and satisfies* (16):

$$KR = GB': \tag{16}$$

C *is of the order* $g \times f$, *as defined by* (17):

$$C = (I_g - KB)F, \tag{17}$$

F *being also of order* $g \times f$ *and satisfying* (12); ω *is a column of* f *orthonormal vectors satisfying* (18):

$$C\omega\xi' = 0, \qquad \omega\omega' = I_f; \tag{18}$$

and θ *is a column of* g *null vectors satisfying:*

$$B\theta = 0, \qquad \theta\theta' = 0. \tag{19}$$

The proof revolves around the properties of K and the products $BK\xi$ and BC.

That a solution K to (16) always exists when R is non-singular is obvious, for then K is uniquely determined as

$$K = GB'R^{-1} \quad (|R| > 0) \tag{20}$$

When R is singular, we can lean on the theory of least-squares to see that there is not only one solution K to (16), but infinitely many. For Lemma 1 assures us that at least one ϕ exists that satisfies (11). If we expand the left number of (21) and notice from (11) that $\xi\phi' = B\phi\phi' = BG$, then we see that (21) is equivalent to (16):

$$(\phi - K\xi)\xi' = 0. \tag{21}$$

For function space, (21) is precisely the set of normal equations for the least-squares predictions of the variables in ϕ from the variables in ξ, and it is well known that a best set of regression weights K always exists, although not uniquely determined when $\xi\xi'$ is singular [13, pp. 151 f.]. The proof in [13] is given as for a function space, but clearly holds for any Euclidean vector space.

Equality (21) reveals the 'statistical' meaning of the proposed construction of ϕ in (15). $K\xi$ is the 'least-squares' estimate of ϕ from ξ, and we have to show that $C\omega + \theta$ is always the matrix of errors of estimate.

That an ω always exists to satisfy (18) should be evident. Indeed, for the first part of (18) it is sufficient to have $\omega\xi' = 0$. When our vectors are all statistical variables, ω can be defined by throwing dice or turning a roulette wheel and making the variances all equal to 1. Or ω can be taken as an orthonormalization of any set of f linearly independent scores that are uncorrelated with those of ξ.

That a θ always exists for (19) is obvious. For example, choose $\theta = 0$. This is the only choice possible if $B'B$ is non-singular, but there are many other possibilities when $B'B$ is singular, especially in a function space.

Granted that the right member of (15) always exists, we have to show its sufficiency, namely, that it satisfies (11). To this end, it will be helpful to establish Proposition 1.

Proposition 1. *If* E *and* M *are real matrices such that* $MEE' = 0$, *then* $ME = 0$. *Similarly, if* ψ *is a column of vectors such that* $M\psi\psi' = 0$ *and the norm of each vector in* ψ *is positive, then* $M\psi = 0$.

The proof of the first part of the Proposition follows from the identity $(ME)(ME)' = MEE'M'$. The right member certainly vanishes when MEE' does, and in particular its main diagonal does; but the sum of the elements of this main diagonal is the sum of squares of all the elements in ME, or each element in ME must vanish. To prove the second part, we can let $\psi = E_0\omega_0$, where ω_0 is an orthogonal basis for the space of ψ. Then $M\psi\psi' = M_0EE_0' = 0$, whence $ME_0 = 0$ from the first part of the Proposition. Therefore $ME_0\omega_0 = M\psi = 0$, or the second part is established.

Notice the requirement in the second part of Proposition 1 that the norm of each element of ψ be positive. This is to ensure that $M\psi$ should actually equal zero and not just be null.

To proceed with the proof of sufficiency for Lemma 2, let ϕ_1 and ϕ_2 be defined respectively as

$$\phi_1 = K\xi, \qquad \phi_2 = C\omega, \tag{22}$$

so that (15) can be rewritten as

$$\phi = \phi_1 + \phi_2 + \theta. \tag{23}$$

Then ϕ_1 is the ' least-squares ' estimate of ϕ from ξ, and ϕ_2 is our candidate for the error of estimate.

Since K always exists, so does ϕ_1. But unlike K, ϕ_1 is uniquely determined by (16) even when R is singular. For suppose K and K^* are any two solutions to (16). Since $(K-K^*)R = 0$, Proposition 1 assures us that $(K-K^*)\xi = 0$, or $K^*\xi = K\xi = \phi_1$ even although $K \neq K^*$. Unlike ϕ_1, ϕ_2 is quite arbitrary, being dependent on a rather arbitrary ω; indeed ϕ_2 is fixed if and only if $C = 0$. Regardless, we shall now show that always

$$B\phi_1 = \xi, \qquad B\phi_2 = 0. \tag{24}$$

Pre-multiply (16) through by B and recall (10) to see that

$$BKR = R, \tag{25}$$

or $(BK-I_n)R = 0$. Therefore, from Proposition 1 $(BK-I_n)\xi = 0$ or—expanding, and recalling the first part of (22)—the first part of (24) holds. Similarly, if we can show that $B(I_g-KB)G = 0$, then from Proposition 1 and from (17) and (12) it will follow that $BC = 0$. Expanding, we find $B(I_g-KB)G = BG-BKBG = RK'-BKRK' = RK'-RK' = 0$, the third member following from (16) and the fourth member from (25). Therefore $BC = 0$. Certainly then $BC\omega = 0$ for any ω, or the second part of (24) always holds.

The proof of (24) could be a bit shorter for non-singular R, for then $BK = I_n$ from (25). It was the case of singular R that required us to invoke Proposition 1.

We can now pre-multiply (23) through by B and use (24) and (19) to obtain $B\phi = \xi$. This verifies that any ϕ constructed according to (15) must satisfy the first part of (11). Curiously enough, according to (24), ϕ_1 also always satisfies the first part of (11). We must now see about the second part of (11).

From (22) and the first part of (18) it follows that

$$\phi_1\phi_2' = 0. \tag{26}$$

Post-multiplying each member of (23) by its own transpose and using (26) and (19) yield

$$\phi\phi' = \phi_1\phi_1' + \phi_2\phi_2'. \tag{27}$$

Post-multiplying the first part of (22) by its own transpose yields

$$\phi_1\phi_1' = KRK'. \tag{28}$$

Post-multiplying the second part of (22) by its own transpose and remembering (17) and (12) yield

$$\phi_2\phi_2' = CC' = (I_g-KB)G(I_g-B'K'). \tag{29}$$

Post-multiplying (16) through by K' shows that

$$KRK' = GB'K' = KBG, \tag{30}$$

the last member being obtained by taking the transpose of the middle member as is justified by the symmetry of the first member. Expanding the last member of (29) and using (30) and (10) yield

$$CC' = G-KRK'. \tag{31}$$

Therefore, from (27), (28), (29) and (31) we see that the second part of (11) is satisfied, or we have completed proof of the sufficiency of (15) for (11).

By capitalizing on the above algebra, the proof of necessity is now relatively rapid. Let ϕ be any solution to (11). Let ϕ_1 be as defined in (22). As we have seen, ϕ_1 is always uniquely determined by ξ, B, and G and thus is always fixed for our problem. It remains to be shown that $\phi-\phi_1$ is necessarily of the form $C\omega+\theta$.

Post-multiplying (21) through by K' and using the definition of ϕ_1 in (22), and then in turn expanding the resulting left member, show that

$$(\phi-\phi_1)\phi_1' = 0, \quad \phi\phi_1' = \phi_1\phi_1' = \phi_1\phi'. \tag{32}$$

Hence, expansion of the left member of (33) and use of (32) yield

$$(\phi-\phi_1)(\phi-\phi_1)' = \phi\phi'-\phi_1\phi_1' = G-KRK', \tag{33}$$

the last member following by recalling (11) and (28). Then from (33) and (31),

$$(\phi-\phi_1)(\phi-\phi_1)' = CC'. \tag{34}$$

If any main diagonal element of CC' vanishes, so must the entire corresponding row of C. Let c be the number of non-vanishing rows of C, let C_c be the non-vanishing c rows of C, and let $(\phi-\phi')_c$ be the c vectors of $(\phi-\phi_1)$ with positive norms. Then from Lemma 1, there is an orthonormal ω_c such that $(\phi-\phi_1)_c = C_c\omega_c$ [in Lemma 1 replace ξ by $(\phi-\phi_1)_c$, B by C_c, G by I_c, and ϕ by ω_c]. Let θ be a column of vectors which is the same as $\phi-\phi_1$ for the $g-c$ null elements, but with zero elements elsewhere. Thus, θ satisfies the second part of (19). Let ω be any column of f orthonormal vectors containing ω_c and arranged so we can write

$$(\phi-\phi_1) = C\omega+\theta. \tag{35}$$

This can be done, for $C\omega$ contains $C_c\omega_c$ and only zeros elsewhere by virtue of the vanishing rows of C. Hence $\phi-\phi_1$ must be precisely of the form implied for $\phi_2+\theta$ in (23).

One detail not explicitly pointed out yet is that $C\omega$ here necessarily satisfies the first part of (18); but this is precisely what (21) implies. A final detail is that θ must satisfy the first part of (19). This follows by pre-multiplying (35) through by B and using (11), (24), and (22).

The proof is thus complete as to the necessity as well sufficiency of (15) for (11), or Lemma 2 has been established.

8. *The Construction of all Solutions to Problem* 1. Definitions (13) for B and G and notation (14) for ϕ transform Lemma 2 into:

Theorem 2. *A necessary and sufficient condition that* η *and* ζ *satisfy Theorem* 1 *is that they be of the form:*

$$\eta = K_q\xi+P\omega+\theta, \qquad \zeta = K_n\xi-AP\omega-A\theta \tag{36}$$

where K_q *and* K_n *are of orders* $q\times n$ *and* $n\times n$ *respectively and satisfy:*

$$K_qR = LA', \qquad K_nR = U^2; \tag{37}$$

P *is of order* $q\times f$ *and satisfies:*

$$PP' = L-K_qRK_q'; \tag{38}$$

ω *is a column of* f *orthonormal vectors satisfying:*

$$P\omega\xi' = 0, \qquad \omega\omega' = I_f, \tag{39}$$

and θ *is a column of* q *null vectors* $(\theta\theta' = 0)$.

Conditions (37) constitute but a direct restatement of (16), letting K_q and K_n be the first q and last n rows of K respectively. In (38), P is the first q rows of C in (17); the right member of (38) is the Gramian submatrix of the first q rows and columns of the right member of (31). The last n rows of C must then be equal to $-AP$, from the fact that $BC = 0$ and from the definition of B in (13). We have used this dependence of the last n rows of C on the first q in writing the coefficient of ω for ζ in (36). Similarly, the θ in (36) is the first q elements of the θ of (15), while $-A\theta$ in (36) is the last n elements of the θ of (15), according to the first part of (19); thus in Theorem 2, the analogue of the first part of (19) always holds and need not be stated explicitly. Orthonormal ω is the same in (39) as in (18), and clearly the first part of (39) is equivalent to that of (18) from the dependence now of the last n rows of C, namely AP, on the first q, or on P.

Thus Theorem 2 is but a restatement of Lemma 2 for the special case where B and G are of the form (13), and its proof has been completed.

If $P \neq 0$ in (36), clearly infinitely many solutions η can be constructed by (36), and in general also infinitely many solutions ζ, by virtue of the arbitrariness of ω. But it is also clear that the extent to which these solutions can differ among themselves, as measured by the norms of their differences, should be limited by how close the main diagonal elements of PP' are to zero. The precise nature of this limitation will be established next.

9. *The Maximal Difference among Alternative Solutions.* Given that ϕ is one solution to (11), let us seek another solution, ϕ^*, to (11) that is maximally different from ϕ in the sense that the norm of each vector in the difference $\phi-\phi^*$ is a maximum. From (23), we can write $\phi^* = \phi_1+\phi_2^*+\theta^*$, where ϕ_1 is the same as for ϕ, but $\phi_2^*+\theta^*$ is specific to ϕ^*. Then $\phi-\phi^*$ is the same as $\phi_2-\phi_2^*$ except for the null column $\theta-\theta^*$, so our problem is to maximize the norms of the elements of $\phi_2-\phi_2^*$. These norms are the main diagonal elements of

$$(\phi_2-\phi_2^*)(\phi_2-\phi_2^*)' = \phi_2\phi_2'-\phi_2\phi_2^* -\phi_2^*\phi_2'+\phi_2^*\phi_2^{*\prime}. \tag{40}$$

Since $\phi_2\phi_2' = \phi_2=\phi_2^{*\prime}$ CC', only $\phi_2\phi_2^{*\prime}$ and $\phi_2^*\phi_2'$—which have identical corresponding

diagonal elements—can vary with ϕ^* in the right of (40). Our problem reduces to maximizing the main diagonal elements of $-\phi_2\phi_2^{*\prime}$. From Schwarz's inequality [1, p. 190], since ϕ_2 and ϕ_2^* have the same norms for corresponding elements, each main diagonal element of $-\phi_2\phi_2^{*\prime}$ cannot be larger than the corresponding diagonal element of $\phi_2\phi_2'$. Clearly, this maximum is attainable *for all diagonal elements simultaneously* by letting $\phi_2^* = -\phi_2$.

To verify that $-\phi_2$ can actually constitute a column of errors of estimate, notice that if $\phi_2 = C\omega$ in (15), then we can write $\phi_2^* = C(-\omega)$, and $-\omega$ satisfies (18) just as well as ω does. According to Lemma 2, our maximally different ϕ^* is as legitimate a solution to (11) as the given ϕ is.

These results can be summarized as:

Lemma 3. *If ϕ is a solution to* (11) *in Lemma* 1, *then there exists a ϕ^* that is maximally different from ϕ, the maximal squares of the norms of the difference column $\phi-\phi^*$ being the main diagona element of*

$$(\phi-\phi^*)(\phi-\phi^*)' = 4CC', \tag{41}$$

where CC' is as given in (29) *and* (31).

The right number of (41) comes from (29) and the fact that $\phi_2-\phi_2^* = 2\phi_2$. The equality in (41) holds since $\phi-\phi^*$ equals $\phi_2-\phi_2^*$ up to a null column.

Returning to Problem 1, Lemma 3 leads to:

Theorem 3. *If η and ζ are a pair of solutions for Theorem* 1, *then there exists a maximally different pair of solutions, η^* and ζ^*, such that:*

$$(\eta-\eta^*)(\eta-\eta^*)' = 4PP' = 4(L-K_qRK_q') \tag{42}$$

$$(\zeta-\zeta^*)(\zeta-\zeta^*)' = 4APP'A = 4(U^2-K_nRK_n'), \tag{43}$$

where the real matrices of (42) *and* (43) *are as defined in Theorem* 2.

The equalities in (42) and (43) are merely expansions of (41) for case (13), using identities established for Theorem 2.

For our application to statistical variables, PP' is the covariance matrix of the errors of estimate of η from ξ, and $APP'A'$ is the corresponding covariance matrix of the errors of ζ from ξ. According to the main diagonals of (42) and (43), the maximal variance of the difference between two alternative solutions for a given factor—common or deviant—is 4 times the variance of the errors of estimate of that factor from ξ. This would lead one to suspect that ξ will have to afford very close estimates of a given pair η and ξ if all other possible pairs of solutions are not to differ materially from this given one. Such a suspicion is well-founded, as the numerical examples of the next section show.

10. *Numerical Examples of the Difference Possible between Alternative Solutions for the same Factor.* From now on, we shall use the special terminology of our statistical problem, although the formulae all remain appropriate to any Euclidean vector space.

Instead of dealing with variances of differences and variances of errors of estimate, it will be convenient to convert them here into correlation coefficients. (In the terminology of Euclidean vector space, we shall treat now the *cosines* of the angles between vectors.) As is well known, if two variables have equal variances σ^2, and if ρ^* is the correlation between the two variables, then the variance of their difference equals $2\sigma^2(1-\rho^*)$. Furthermore, if one of these variables has multiple correlation ρ on a set of predictors, then the variance of the errors of estimate equals $\sigma^2\rho^2$. Using these two facts in conjunction with the main diagonals of (42) and (43) shows that for each factor, common or deviant, $2\sigma^2(1-\rho^*) = 4\sigma^2\rho^2$, whence

$$\rho^* = 2\rho^2-1, \tag{44}$$

where ρ^* is the correlation between an element y of η (z of ζ) with the corresponding element of the maximally different η^* (ζ^*), and ρ is the multiple correlation of y (z) on ξ.

While ρ^2 in (44) can vary between 0 and 1, ρ^* can vary more widely between -1 and 1. It may be useful to exhibit specific numerical values for the relation (44). These are displayed in Table 1.

Table 1. *The Minimal Correlation* (ρ^*) *always attainable between Two Alternative Solutions for the Same Factor* (*Common or Deviant*), as a *Function of the Multiple Correlation* (ρ) *of that Factor on the Observed Scores.*

ρ	·00	·30	·60	·80	·90	·95	·97	·99
ρ^*	−1·00	−·82	−·28	·28	·62	·81	·88	·96

To ensure any positive correlation at all between y and y^* (z and z^*), the multiple correlation of y (z) on ξ must exceed ·71. But Table 1 reveals that even when ρ reaches ·80, the corresponding ρ^* is only ·28. The situation improves, but not too much, when a ρ of ·90 is reached. Only when a ρ as large as ·99 is attained do the alternative solutions necessarily begin to correlate by as much as ·96 with each other.

In a previously published numerical example where $q = 3$ and $n = 7$, it turned out that the ρ were ·90, ·89, and ·88 respectively for the elements of η, while they ranged from ·59 to ·92 for the elements of ζ [5, pp. 182 f.]. It should now be evident that having determined U^2, A, and L for such data does not go too far towards fixing any corresponding score matrices η and ζ.

Godfrey Thomson reported that the ρ for nine ' primary traits ' in Thurstone's original analysis range from ·630 to ·908, and remarked that: ' Those correlations do not look so bad ' [15, p. 339]. If we look at these ρ again from the point of view of ρ^*, it seems that the sought-for traits are not very distinguishable from radically different possible alternative traits for the identical factor loadings.

11. *The Case of Non-Singular R.* When R is non-singular, some formulae above can be conveniently expressed in terms of R^{-1}. From (37), we have K_q and K_n uniquely defined by

$$K_q = LA'R^{-1}, \qquad K_n = U^2R^{-1}. \tag{45}$$

If we let η_1 and ζ_1 be the predicted values of η and ζ respectively from ξ, and η_2 and ζ_2 the corresponding errors of prediction up to a null column, then from (36) and (45) we can write:

$$\eta_1 = LA'R^{-1}\xi, \qquad \zeta_1 = U^2R^{-1}\xi, \tag{46}$$

$$\eta_2 = P\omega, \qquad \zeta_2 = -AP\omega. \tag{47}$$

From (46) and (4), we find

$$\eta_1\eta_1' = LA'R^{-1}AL, \qquad \zeta_1\zeta_1' = U^2R^{-1}U^2. \tag{48}$$

Now, from (27),

$$\eta\eta' = \eta_1\eta_1' + \eta_2\eta_2'. \tag{49}$$

Then, from (49) and the first parts of (7), (47), and (48), we have the identity:

$$\eta_2\eta' = PP' = L - LA'R^{-1}AL. \tag{50}$$

Similarly, since

$$\zeta\zeta' = \zeta_1\zeta_1' + \zeta_2\zeta_2', \tag{51}$$

we have the identity

$$\zeta_2\zeta_2' = APP'A' = U^2 - U^2R^{-1}U^2. \tag{52}$$

An interesting result of (52) is that, if U^2 is non-singular, we derive an identity for R^{-1}:

$$R^{-1} = U^{-2} - U^{-2}APP'A'U^{-2}. \tag{53}$$

This is obtained by pre- and post-multiplying (52) through by U^{-2}.

12. *The Case where L and* U^2 *are Non-Singular: The Matrix Q.* If both L and U^2 are non-singular, further interesting and useful special identities emerge. First, let us notice the important fact that U^2 non-singular implies R non-singular, which may be worth stating as a theorem:

Theorem 3. *If R,* U^2, *and* $R - U^2$ *are each Gramian, and if* U^2 *is non-singular, then R is non-singular.*

This follows from the theorem that the rank of sum of two Gramian matrices cannot be smaller than the rank of either term in the sum [4, p. 73], and clearly $R = (R - U^2) + U^2$. In our case, $R - U^2 = ALA'$, according to (2).

Notice that L non-singular by itself says nothing at all about the rank of R, but the rank of ALA' does (being a lower bound to the rank of R).

Since we now assume both L and U^2 to possess inverses, we are at liberty to define a real symmetric matrix Q of order q by:

$$Q = L^{-1}+A'U^{-2}A. \tag{54}$$

With this definition, we immediately establish the following theorem:

Theorem 4. *The matrix Q as defined in* (54) *is Gramian and non-singular. Furthermore,*

$$Q^{-1} = PP' = L-LA'R^{-1}AL, \tag{55}$$

where P is as defined in (38), *or Q^{-1} is the covariance matrix of the estimates of the common-factor scores from ξ.*

Perhaps as easy a way as any to prove Theorem 4 is to multiply the right member of (54) by the last member of (55) and see that the product (in either order) reduces to Iq, using the fact that $ALA' = R-U^2$.

That the inverse of R in the right of (55) exists is assured by Theorem 3. Indeed, we can re-write (53) as:

$$R^{-1} = U^{-2}-U^{-2}AQ^{-1}A'U^{-2}. \tag{56}$$

Identity (56) was established before in [3, pp. 91 f.] for the special case of the δ-law (U^2 diagonal) and where $L = I_q$ and $q = r$. For that case, it often serves as an economical way of inverting R by inverting a usually much smaller matrix Q [5]. The identity holds more generally for the β-law, for $L \neq I_q$, and for $q \neq r$, although it need not lead to parsimonious calculations in the general case.

Pre-multiplying (56) through by LA' and recalling (45) and (54) establishes a further useful identity:

$$K_q = LA'R^{-1} = Q^{-1}A'U^{-2}. \tag{57}$$

For the δ-law and assuming $L = I_q$ and $q = r$, (57) was established previously in [3] and [14], being generalized to $L \neq I_r$ in [12, p. 279]. We have now further generalized it to the β-law and $q \neq r$.

Another way of arriving at (57) is to note from (26) and (14) that

$$\eta_1\zeta_1'+\eta_2\zeta_2' = 0. \tag{58}$$

Hence, from (58), (45), (46), and (47),

$$K_qU^2 = LA'R^{-1}U^2 = PP'A'. \tag{59}$$

For (59), only the assumption that R^{-1} exists is needed. If in addition both U^2 and L are non-singular, we can immediately derive (57) from (59), using (55).

The error covariance matrices can now be written as

$$\eta_2\eta' = Q^{-1}, \qquad \zeta_2\zeta_2' = AQ^{-1}A', \tag{60}$$

according to (50), (52), and (55). No diagonal element of Q^{-1} can vanish from the Gramian nature of Q^{-1}. Similarly, at least r diagonal elements of $AQ^{-1}A'$ must be positive, for the rank of $AQ^{-1}A'$ is the rank of A, again from the Gramian nature of Q^{-1} [4, p. 75]. Since these diagonal elements are the *variances* of the errors of prediction from ξ, we can state:

Theorem 5. *For finite n, if both L and U^2 are non-singular, all common-factors and at least r deviant-factors have positive variances of errors of prediction from ξ.*

For the sake of completeness, we may remark that identities for the covariance matrices of the predictions are, respectively:

$$\eta_1\eta_1' = L-Q^{-1}, \qquad \zeta_1\zeta_1' = U^2-AQ^{-1}A', \tag{61}$$

as follows from (49), (51), (7), and (60).

13. *The Infinite Universe of Content and the Infinite Population.* A multiple correlation coefficient cannot decrease as the number of predictors increases: in general the coefficient increases. Hence, as n increases, it may be expected that the multiple correlation ρ on ξ will increase in general for each element of η and ζ. Although for any finite n, and when no main diagonal element vanishes in PP' or $APP'A'$, none of the correlations ρ can equal unity, this does not prevent their *limits* from being unity as $n \to \infty$. In particular, though L and U^2 may be non-singular for all n, so that at least $q+r$ factor scores are not perfectly predictable for any n according to Theorem 5, nevertheless *all* factor scores can possibly have their respective $\rho \to 1$ as $n \to \infty$.

In developing a psychological theory, any set of n observed variables to be factored will ordinarily be regarded as but a sample from an infinitely large universe of content. Since we now wish to study what may happen as $n \to \infty$, we can no longer in general consider N to be possibly finite. Of special interest is where R is non-singular for all n. In order for R to be non-singular in (4), clearly it is necessary that N be not less than n. (Indeed, if the linear restriction is laid on the x_{ji} that their means over i be zero, we must have $N-1 \geqslant n$.)

In any event, finite N is not the case of greatest interest. Not only does this restrict n to being finite for non-singular R, but it prevents the x_j from being *continuous* random variables (in particular, from having a theoretical normal distribution). In a quest for scientifically meaningful factors, finite N occurs usually for but a *sample* from an infinite population. Sampling problems cannot be studied profitably until the population problems are defined. Sampling considerations are secondary to the basic problems with which we are concerned here. Our problems remain even if there is no sampling error due to people, and do not arise because of such error.

From now on we hypothesize that $N = \infty$; our basic population may be countably or not countably infinite. It was largely for this purpose that the conventions about score matrices were introduced in § 3 above. We now wish to consider what may happen as n increases.

In the general case, q and r are functions of n, so we shall write q_n and r_n ($n = 1, 2,...$). If U^2, A, and L are determined for a given n, the elements of these matrices may possibly have to be modified—apart from the enlargement of their orders with n, q_n, and r_n—to keep $R-U^2$ Gramian as n increases. (In the special case where q_n and r_n can remain constant for all n sufficiently large, say for all $n \geqslant n'$, then the L determined for $n = n'$ can possibly remain constant thereafter, as can also the first n' rows and columns of U^2 and the elements in the first n' rows of A.)

To emphasize their dependence on n, we shall now attach the subscript n also to the matrices defined previously in this paper. In the case of the column score matrix ξ_n of the first n observed variables, the dependence on n takes the form only of enlarging the number of variables in the matrix and not in changing those already in the matrix for a smaller value of n. Similarly, the observed intercorrelation matrix R_n depends on n only in that the number of rows and columns increases with n; but once an element occurs for a given n, it stays fixed in value as n increases. Not so for R_n^{-1}; here each element changes value in general as n increases [cf. 8, p. 281]. In consequence, each element of K_{qn} and K_{nn} in (45) may change with n, as well as of P_n in (50) and (52). Correspondingly, η_n and ζ_n may have each element vary as n increases, according to the construction in (36). In particular, each of the elements of each of the column factor components may vary: η_{1n}, η_{2n}, ζ_{1n}, ζ_{2n}.

14. *Conditions for Determinacy.* What we have to consider in place of (2) is a *sequence* of equalities:

$$R_n = A_n L_n A_n^2 + U_n^2 \qquad (n = 1, 2,...) \tag{62}$$

for which we seek columns η_n and ζ_n such that, in place of (1),

$$\xi_n = A_n \eta_n + \zeta_n \qquad (n = 1, 2,...) \tag{63}$$

and (6) and (7) are satisfied for each n (or better, for all n sufficiently large).

We shall say that a factor score (element of a column score matrix) associated with the sequence (62) is *determinate* if and only if $\lim_{n\to\infty} \rho_n = 1$ for this factor. Clearly, a necessary and sufficient condition for the kth common-factor to be determinate is that the kth diagonal element of $\eta_{2n}\eta'_{2n} = P_n P'_n$ tend to zero as $n \to \infty$. Similarly, a necessary and sufficient condition for the jth deviant-factor to be determinate is that the jth diagonal element of $\zeta_{2n}\zeta'_{2n} = A_n P_n P'_n A'$ tend to zero as $n \to \infty$. But the vanishing of the diagonal element of a Gramian matrix necessitates the vanishing of the entire corresponding row and column. The vanishing of all diagonal elements implies the vanishing of the entire Gramian matrix. Hence, the theorem:

Theorem 6. *A necessary and sufficient condition that all common-factors be determinate for* (62) *is that the left member of* (64) *exist and* (64) *hold:*

$$\lim_{n\to\infty} P_n P'_n = 0. \tag{64}$$

The corresponding condition for deviant-factors is that the left member of (65) *exist and* (65) *hold:*

$$\lim_{n\to\infty} A_n P_n P_n' A_n' = 0. \tag{65}$$

If R_n *is non-singular for all* n *sufficiently large, then* (64) *is equivalent to*

$$\lim_{n\to\infty} (L_n - L_n A_n' R_n^{-1} A_n L_n) = 0, \tag{66}$$

and (65) *is equivalent to*

$$\lim_{n\to\infty} (U_n^2 - U_n^2 R_n^{-1} U_n^2) = 0. \tag{67}$$

If L_n *and* U_n^2 *are non-singular for all* n *sufficiently large, then* (64) *and* (65) *are equivalent to the respective parts of* (68)*:*

$$\lim_{n\to\infty} Q_n^{-1} = 0, \qquad \lim_{n\to\infty} A_n Q_n^{-1} A_n' = 0, \tag{68}$$

where Q_n *is as defined in* (54) *for all* n *sufficiently large.*

The equivalence of the above conditions to each other follows from (50), (52), and (60).

Evidently, from (44), if $\rho_n \to 1$ then $\rho_n^* \to 1$ for the same factor. Or, from (42) and (43), if (68) holds then all variances of differences between alternative solutions tend to zero as $n \to \infty$. Hence Theorem 7:

Theorem 7. *A sequence* (62) *can lead to at most one column of determinate common-factors, say* η_∞*, and at most one column of determinate deviant-factors, say* ζ_∞*. That is, all alternative solutions for a determinate factor are equal up to null columns.*

15. *The Case of Diagonal* U_n^2 (*the* δ*-law*). Although we have used the terms 'common' and 'deviant' for the η_n and ζ_n, there is nothing basic in the algebra thus far that distinguishes between the two types of factors. The β-law as expressed by (6) is essentially symmetric in η and ζ. Indeed, our algebra would be symmetric throughout if we had considered not η but $A\eta$ in (1) or (63). Essential asymmetry begins to occur when further restrictions beyond the β-law are laid on either η or ζ. The δ-law, or combining the γ-law with the β-law, leads to a particularly fundamental type of asymmetry, for then diagonal U_n^2 becomes intimately related to the main diagonal of R_n^{-1}.

Let u_{jn}^2 be the jth diagonal element of U_n^2. If u_{jn}^2 is always positive, then clearly in order for (67) to hold the non-diagonal elements in the jth row and column of R_n^{-1} must tend to vanish as n increases. This implies:

Theorem 8. *If all deviant-factors are determinate and if* U_n^2 *is diagonal and non-singular for all* n *sufficiently large, then* R_n^{-1} *must tend to a diagonal matrix as* $n \to \infty$.

Notice that Theorem 8 assumes U_n to be non-singular for all n sufficiently large, so R_n^{-1} exists for all n sufficiently large according to Theorem 3. That the main diagonal of R_n^{-1} must always have a limit as $n \to \infty$ is a consequence of multiple-correlation theory, as will be marked in more detail in the next section below.

The tendency of R_n^{-1} to a diagonal matrix has been discussed from a point of view closely related to the present one in [3, p. 95]. As has been pointed out in [8, p. 282 and pp. 294 f.), this tendency can serve as a basis for a test of the hypothesis that there is a useful δ-law structure at all to the universe of content. If R_n^{-1} does *not* tend to a diagonal matrix, then there can be no determinate unique-factor scores, nor hence a determinate common-factor space.

Our condition for the non-diagonal elements of R_n^{-1} to tend to vanish is somewhat different from—though very closely related to—that in [3]. We require in Theorem 8 only that the unique-factors be determinate and have non-vanishing uniqueness for all n sufficiently large.

III. SUPPLEMENTARY PROBLEMS

16. *A Solution to Problem II* (*Communalities*). By considering only what happens to the main diagonal of R_n^{-1} as n increases, we arrive at a solution to the problem of communalities of the δ-law. As is well known, the jth main diagonal element of R_n^{-1} is the reciprocal

of the variance of the errors of estimating x_j from the remaining $n-1$ variables from the observed universe. Let σ^2_{jn} denote this variance [the 'partial anti-norm' of x_j in the terminology of (8)].

Inspecting the main diagonals of (67) shows that for each j:

$$\lim_{n\to\infty} u^2_{jn}\left(1-\frac{u^2_{jn}}{\sigma^2_{jn}}\right) = 0. \tag{69}$$

Since a variance of errors of estimate cannot increase as the number of predictors increases, σ^2_{jn} is a monotonely decreasing function of n for all j, so there always exists $\lim_{n\to\infty} \sigma^2_{jn}$ (the 'total anti-norm' of x_j), which we shall denote by $\sigma^2_{j\infty}$. (We have used this existence proposition in Theorem 8; it implies that each element of the main diagonal of R_n^{-1} always tends to a limit, finite or infinite, as $n \to \infty$.)

Considering (69) and its source in (52), we can state:

Theorem 9. *If there exists* $\lim_{n\to\infty} u^2_{jn}$, *say* $u^2_{j\infty}$, *then a necessary and sufficient condition that the jth unique-factor associated with sequence* (62) *be determinate for diagonal* U^2_∞ *is that either* $u^2_{j\infty} = 0$ *or* $u^2_{j\infty} = \sigma^2_{j\infty}$.

The restriction of R_n to being non-singular has been omitted in Theorem 9, for it has been shown in [8, p. 293] that $u^2_{jn} \leqslant \sigma^2_{jn}$ even when R_n is singular. [That $u^2_{jn} \leqslant \sigma^2_{jn}$ when R_n is non-singular follows also from (52) for diagonal U^2_n.] Hence, if $\sigma^2_{jn} = 0$ for any n, $u^2_{jn} = 0$ for that n, and Theorem 9 holds also for this singular case.

According to Theorem 9, if the main diagonal of R_n is to be modified by subtraction of a non-singular U^2_n to leave only the contribution of common-factors in (62), the choice is limited only to a U^2_n which will tend to whatever the main diagonal of R_n^{-1} tends. This provides a solution to the problem of communalities. *From the point of view of determinacy of factors, the ideal choice of uniquenesses is* $\sigma^2_{j\infty}(j = 1, 2,...)$.

A striking feature of this solution for uniquenesses (or equivalently, for communalities) is that it lays no restrictions on the common-factor space. Nothing is said about the *rank* of $R_n - U^2_n$ for any n, nor about the determinacy or location of the common-factors. The limit of the rank of the common-factor space can be finite or infinite. This may help explain why past attempts to solve the problem of communalities for finite n have not been successful. Apparently, only lower bounds to communalities are universally possible for finite n [cf. 10]. The ρ^2_{jn} afford such bounds, and must tend to the actual communalities if the unique-factors are to be determinate with positive uniquenesses.

17. *Implications for Problem III: the Scientific Meaning of Factor Analysis.* Since Spearman's original analysis of the problem when $q_n = r_n = 1$ and U^2_n non-singular ($n = 1, 2,...$), it has been generally recognized that factor scores cannot be estimated exactly from the observed content scores.

For example, Holzinger and Harman state: "Since the total number of factors (both common and unique) exceeds the number of observed variables, the value of any particular factor for a given individual cannot be obtained by direct solution but can only be estimated from the observed values of the variables. The best prediction, in the least-square sense, is that obtained by the ordinary regression method" [12, p. 264]. Thurstone seems to regard this estimation problem as largely a practical one; and his position is that "the problem of appraising an individual as to each of several factors is best solved in the practical situation by using a test composite for each factor" [16, p. 515].

These authors are focusing on a common-factor score matrix η_n. They apparently take it for granted not only that η_n exists, but also that it is uniquely determined by U^2_n, A_n, and L_n, even though it can be only approximated as a linear function of the variables in ξ_n. They do not discuss what can be said about η_n beyond its predictability from ξ_n.

Thomson has gone more intensively into this problem, and Ledermann has generalized his conclusions. Assuming one pair of solutions η_n and ζ_n exists, they show how further solutions can be obtained from this pair for the same A_n and U^2_n, assuming $L_n = I_{r_n}$ [cf. 15, pp. 371 f.]. While recognizing that "usually the accuracy is very low indeed" for predicting factors from the observed ξ_n [15, p. 336], Thomson has not considered the relationship of ρ^* to ρ as was done in §§ 9-10 above. As remarked above, Thomson considered that ρ between ·630 and ·908 "do not look so bad", whereas from (44) this implies ρ^* between −·206 and ·650.

It seems safe to say that there has been considerable complacency about being able to ascribe particular scientific meaning to η_n and ζ_n on the basis of A_n, L_n, and U^2_n, even if the

factors are not determined closely for the given n. It has been especially true with respect to common-factors that they have been *named* according to the content of the observed variables that have ' high ' loadings on them in A_n.

Thurstone has further insisted that " *it is a fundamental criterion of a valid method of isolating primary abilities that the weights of the primary abilities for a test must remain invariant when it is moved from one test battery to another test battery* " [16, p. 361, italics his]. Here again is expressed the belief that A_n delimits some particular η_n with no reference to the ρ_n or ρ_n^*.

It appears from the relation of ρ_n^* to ρ_n (for each intended factor) that the predictability of factors from ξ_n is not merely a practical problem. If ρ_n^* is low, it raises the question of what it is that is being estimated in the first place; instead of only *one* ' primary trait ' there are many widely different variables associated with a given profile of loadings. If the scientific usefulness of factor analysis is to rest on its finding factor scores that can serve as meaningful reference axes, surely these reference axes should be defined by the analysis as being distinct from alternative axes.

No matter how well an A_n satisfies a given criterion of rotation, and no matter how constant some of its rows may be from test battery to test battery, a particular set of meaningful reference axes is still not defined if ρ_n does not tend to unity for these axes.

If in practice ρ_n does not get to be well above ·90, say, for most of the factors, so that ρ_n^* is generally above ·60, it is hard to see that any meaningful reference axes have been established by computing A_n, L_n, and U_n^2, regardless of what arithmetical criteria and manipulations are used.

Attention may be focused on the problem whether *in principle* a given universe of data admits of an approximately determinate set of common- and deviant-factors. Empirical evidence is scanty in published studies, since the ρ_n are not usually computed. The few instances of recorded ρ_n noted above are not very encouraging for the hypothesis of determinacy. The new radex approach to factor analysis [9] does not require determinacy in order for practical and theoretical use to be made of its analysis of an R_n. Not so the Spearman-Thurstone type of theory, which for mental tests is in effect a neo-faculty theory of psychology. Unless the factors or ' faculties ' are pinned down, they can be used neither in theory nor in practice; hence, if determinacy is unobtainable from the observed universe of ξ, a correlational analysis of R_n by itself is not adequate for the Spearman-Thurstone type of theory. A more direct and experimental approach to the hypothesized factors may be required to test and verify their nature. If more direct observations on the η_n and ξ_n cannot be made than statistical analysis of R_n and ξ_n, the Spearman-Thurstone approach may have to be discarded for lack of determinacy of its factor scores.

18. *Problem IV: ' Inverted ' Factor Analysis.* The proposal has often been made to ' factor ' people rather than variables, especially when N is finite and smaller than n [cf. 15, chaps. XIII and XIV]. Discussing such a proposal only for finite N and n clearly ignores the problem of determinacy of factor scores. Artifacts due to finiteness of the sample data certainly occur in the algebra of the usual discussions. The variables can no longer theoretically have normal distributions, as already remarked, nor any other type of continuous distribution.

There is nothing mathematically wrong with the ' inverted ' problem of factoring people, provided the proper framework is specified. For determinacy, in general n must be infinite. One could consider what happens as $N \to \infty$ analogously to our present treatment. This requires specifying, somehow, inner-products for the Euclidean vector space implied, the vectors corresponding to people now rather than to variables. How to specify these seems to be as yet an unsolved empirical problem.

The most general approach may ultimately be not to ' factor ' either people or variables, but to start out by assuming both N and n infinite, and to regard the totality of theoretically possible observations as defining a *Hilbert space*, rather than a finite-dimensional vector space. Our present theory started with finite n, and considered what happened as $n \to \infty$. This implicitly assumes some kind of frequency distribution of variables *over people*. A related concept of a frequency function for variables occurs explicitly in the simplex subtheory of the radex approach to factor analysis [11]. A more general formulation is under way by the writer in terms of Hilbert space and measure theory that may simultaneously take care of direct and inverted factor theory in a manner that seems natural for the problem of factor analysis, and avoiding the artifacts and indeterminacies of finite data.

Neither N nor n should ordinarily be regarded as finite at the outset in a fundamental theory for mental tests, say, or for other substantive areas for factor analysis. Starting with finite n in our present theory, or for other substantive areas of our present theory. Starting with a finite N present paper may be regarded as an incompleteness of incompleteness, apart from the other problems for an ' inverted ' factor theory is an analogous sign of incompleteness, apart from the other problems involved.

19. *Problem V: Second-Order Common-Factors.* The solution to Problem I has important implications for the concept of ' second-order ' factors. It has been suggested that when $q = r$ and $L \neq I_r$, or oblique common-factors are used, it may sometimes be meaningful to ' factor ' these r common-factors in turn [16, chap. XVIII]. The procedure would be to seek a diagonal matrix, say D, such that $L - D^2$ would be Gramian and of rank less than r; say the rank would be s where $s < r$. The assumption is that thereby s ' second-order ' common-factors and r ' second-order ' unique-factors are determined which underly the original r common-factor.

Now, Problem I applies to the second-order factors as well as to the first-order ones. If $|D^2| > 0$, then the solution to Problem I shows that infinitely many sets of second-order factors will satisfy any given rotation in this ' second-order ' space. Only as $r \to \infty$ can the embarrassment of too many solutions possibly disappear.

However, advocates of second-order factors usually also advocate having r finite and relatively small. Requiring r to become indefinitely large would contradict a basic hypothesis of this school of thought. If r is to be finite, this requires either: (*a*) giving up the notion of second-order factors, or (*b*) hypothesizing that the second-order factors scores are *in principle indeterminate* for any given rotation of axes in the second-order space. In case (*b*), the question might be raised as to the psychological meaning or usefulness of the concept of second-order factors as reference axes.

20. *Problem VI: Rotation of Axes.* One of the currently outstanding questions of common-factor theory is that of rotation of axes. This concerns the term $A\eta$ in (1). Given that A and η satisfy (1), let T be any real non-singular matrix of order q and let $\overline{A}$, $\overline{\eta}$, and $\overline{L}$ be defined by

$$\overline{A} = AT^{-1}, \qquad \overline{\eta} = T\eta, \qquad \overline{L} = TLT'. \tag{70}$$

Then clearly

$$\overline{A}\overline{\eta} = A\eta, \tag{71}$$

or $\overline{A}$ and $\overline{\eta}$ satisfy (1). Furthermore, $\overline{\eta}$ satisfies (6), and also (7) and (9) with $\overline{A}$ and $\overline{L}$ in place of A and L.

$\overline{\eta}$ is called a ' rotation ' of η, being an ' orthogonal rotation ' if T is an orthogonal matrix. $\overline{\eta}$ is a solution to Problem I when A and L are replaced by $\overline{A}$ and $\overline{L}$. Hence there are as many $\overline{\eta}$ corresponding to a given rotation matrix T as there are η for (1) as is. The same indeterminacy problem of scores holds for all rotations if only an indirect analysis is made via A and T. Since the criteria for rotations used by different schools of thought are in terms of the parameters or elements of $\overline{A}$, these again ignore the determinacy problem of scores. Although we arrived at a solution to Problem II, or the determination of U^2, this does not help at all to solve Problem VI or the fixing of T when A is given.

Enlarging the perspective of Problem VI to include also Problem I suggests that perhaps undue emphasis has been placed on Problem VI in the past. If reference scores $\overline{\eta}$ are not going to be pinned down thereby anyway, according to the solution to Problem I, why bother about rules for $\overline{A}$ and T? Such rules alone may not enable one to make much use of the implied factors, or to distinguish them from quite different factors.

As remarked in § 17, perhaps factor analysis has strained too much to try to isolate meaningful factors by means of a correlational analysis alone. Many important and useful things can be learned about the structure of ξ and R from a correlational analysis alone, without recourse to reference factor scores, as radex theory has shown [9, 11]. Perhaps we should be content with this latter type of structural information until we can make direct observations and experiments beyond ξ and R in order to determine $\overline{\eta}$.

REFERENCES

1. Birkhoff, Garrett, and MacLane, Saunders (1953). *A Survey of Modern Algebra* (rev. ed.). New York: Macmillan.
2. Dwyer, P. S. (1940). 'The evaluation of multiple and partial correlation coefficients from the factorial matrix.' *Psychometrika*, V, 211-232.
3. Guttman, Louis (1940). 'Multiple rectilinear prediction and the resolution into components.' *Psychometrika*, V, 75-99.
4. Guttman, Louis (1942). 'Properties of Gramian matrices', Chap. II, *The Prediction of Quantitative Variates by Factor Analysis* (Ph.D. Thesis, Univ. Minnesota).
5. Guttman, Louis, and Cohen, Jozef (1943). 'Multiple rectilinear prediction and the resolution into components: II.' *Psychometrika*, VIII, 169-183.
6. Guttman, Louis (1944). 'General theory and methods for matric factoring.' *Psychometrika*, IX, 1-16.
7. Guttman, Louis (1952). 'Multiple group methods for common-factor analysis: their basis, computation, and interpretation.' *Psychometrika*, XVII, 209-222.
8. Guttman, Louis (1953). 'Image theory for the structure of quantitative variates.' *Psychometrika*, XVIII, 277-296.
9. Guttman, Louis (1954). 'A new approach to factor analysis: the radex', in Lazarsfeld, Paul F. (ed.), *Mathematical Thinking in the Social Sciences*. Glencoe, Ill.: The Free Press.
10. Guttman, Louis (1954). 'Some necessary conditions for common-factor analysis.' *Psychometrika* XIX, 149-161.
11. Guttman, Louis (1955). 'A generalized simplex for factor analysis.' *Psychometrika*, XX: (in press).
12. Holzinger, K. J., and Harman, H. H. (1941). *Factor Analysis*. Chicago: Univ. Chicago Press.
13. Jackson, D. H. (1930). 'The theory of approximation.' *American Mathematical Society Colloquium Publications*, II. New York: Am. Math. Society.
14. Ledermann, W. (1939). 'On a shortened method of estimation of mental factors by regression.' *Psychometrika*, IV, 109-116.
15. Thomson, G. H. (1950). *The Factorial Analysis of Human Ability*. London: Univ. London Press.
16. Thurstone, L. L. (1947). *Multiple Factor Analysis*. Chicago: Univ. Chicago Press.

ERRATA

The editor regrets that the following errors have been made in printing Professor Guttman's article on 'An Additive Metric from all the Principal Components of a Perfect Scale' in the preceding issue of this *Journal*, (VIII, pp. 17-24).

P. 22, equation (3). *Instead of* "δ" *read* "δ_{ij}".

P. 23, equation (15). *The right member should read :* $-\frac{1}{\eta_a^2} f_0 x_{a0}$.

P. 23, equation (16). *In the right member instead of* "$x_{ai,\,i+1}$" *read* "$x_{a,\,i+1}$".

P. 23, equation (20). *Instead of* "p_{ij}" *read* "p_{0j}".

P. 24, equation (22). *Instead of* "p_{ij}" *read* "p_{1j}".

Vol. VIII Part II The British Journal of Statistical Psychology November 1955

FACTOR ROTATION FOR PROPORTIONAL PROFILES: ANALYTICAL SOLUTION AND AN EXAMPLE

By R. B. CATTELL and A. K. S. CATTELL
Laboratory of Personality Assessment and Group Behaviour, University of Illinois

I. *Existing Attempts at Unique Factor Resolution.* II. *The Rationale of Proportional Profiles.* III. *History of Empirical Explorations of the Problem.* IV. *Analytical Solution for Proportional Profiles with Orthogonal Axes.* V. *A Worked Example.* VI. *Summary.*

I. EXISTING ATTEMPTS AT UNIQUE FACTOR RESOLUTION

Factor analysts may be divided—rather sadly divided—into (*a*) those for whom factors are mainly mathematical utilities, specific in reference to a particular matrix or mathematical process and quite restricted in psychological meaning (if any), and (*b*) those for whom factor analysis is a means of discovering and dealing with scientific concepts that persist through all matrices concerned with similar phenomena. With increasing realization of the scope of factor analysis in the social sciences, and with the accumulation of invariant results [8], the emphasis on factors as scientific hypothesis has fortunately begun to prevail. The need to expand this conceptual usage of factors makes it all the more urgent to discover a technique for determining in any factorial investigation that unique rotational solution which will correspond to the presumed inherent structure in nature.

Some seven distinct principles in use for gaining a unique, meaningful solution were listed in an earlier article [5]; six of them have been enlarged upon in a book [7] which also reviewed critically the principle of criterion rotation [12], but not the proposal of Sandler [21] which, though ingenious, merely passes the responsibility from *R*- to *Q*-technique and asks the blind to lead the blind. Of existing methods none seems to the writer to have been so successful as ' simple structure '; for, in the first place, it has a reasonably clear theoretical basis—the principle of parsimony; and, secondly, it has proved itself pragmatically, that is, so far as pragmatism can ever be acceptable. The pragmatic proof consists in having yielded psychologically meaningful factors more often than other methods have done, and in providing an impressive accumulation of comparisons [8, 13, 15] that have revealed invariance of factor pattern from one study to another—a criterion less tainted by subjectivity than the first.

It would, however, be a mistake to assume that simple structure can be accepted as a final satisfactory solution to the rotation problem. It suffers from at least three shortcomings: (1) different users do not define and use it according to a single acceptable definition; (2) the labour and skill required to obtain simple structure is such that half the people who set out to obtain it fall far short of doing so; and (3) the procedure has certain theoretical weaknesses. The second difficulty promises to be removed by electronic computers, thanks to the methods devised by Carroll [3], Neuhaus and Wrigley (Quartimax) [19], Saunders [22], Thurstone [26] and Tucker [28], though recent experience shows that they are not altogether fool-proof and may still require to be supplemented by craftsmanship. However, it is the theoretical inadequacies of the procedure that are important and will concern us here.

The main irremediable theoretical failure arises from the fact that, if a factor is to be located by simple structure, it must leave an appreciable fraction of the variables unloaded, i.e., in a discernible hyperplane. Now it is conceivable, and indeed it often happens, that a factor will affect *all* the variables in a matrix. This can occur first because (*a*) the variables have not been chosen with sufficient catholicity and with due regard for ensuring ' hyper-plane

stuff' [7], e.g., by using the 'personality sphere' concept [6] or some similar principle of variable sampling: this is an avoidable fault of design, though judging by its frequency it is hard to escape. Secondly, it may occur because (*b*) there exist factors which are indeed so pervasive and ubiquitous (like *g* in the cognitive field, or surgency-desurgency in the field of personality) that no clear hyperplane can ever be found for them. The converse but equally disconcerting finding is that in one and the same factorization, with *n* factors extracted, one can sometimes find more than *n* hyperplanes, so that there exist alternative resolutions. The senior author has found and reported such equally attractive alternatives [8, 10], though in a prolonged experience they have been rare. Other writers, e.g., Bargmann [1], maintain they are quite frequent; and, judging from the occurrence of alternative solutions by competitive psychologists, the discrimination of the better from the poorer or false hyperplane is not always easy. The principle now to be propounded considers real structure to be inherent in the data, but does not depend for resolution on the elusive hyperplane.

II. THE RATIONALE OF PROPORTIONAL PROFILES

The principle called 'parallel proportional profiles' was proposed in 1944 [3] as a method for avoiding the restrictions of simple structure. Only a brief indication of its nature therefore need be given here.

Since the assumptions and objectives of proportional profiles are intrinsically difficult to convey, as evidenced by the number of misunderstandings, certain points must be emphasized. The basic assumption is that, *if a factor corresponds to some real organic unity, then from one study to another it will retain its pattern, simultaneously raising or lowering all its loadings according to the magnitude of the role of that factor under the different experimental conditions of the second study.* No inorganic factor, a mere mathematical abstraction, would behave in this way, within the realm of orthogonal factors. The principle suggests that every factor analytic investigation should be carried out on at least two samples, under conditions differing in the *extent* to which the same psychological factors (working as independent, orthogonal influences) might be expected to be involved. We could then anticipate finding the 'true' factors by locating the unique rotational position (simultaneously in both studies) in which each factor in the first study is found to have loadings which are proportional to (or are some simple function of) those in the second: that is to say, a position should be discoverable in which the factor in the second study will have a pattern which is the same as in the first, but is stepped up or down. Obviously, if no change in the relative importance of the same factors occurs from the first to the second experiment, no rotational position can be determined; but it is assumed that, either by natural influences or by deliberate experimental control, the real factor influences will differ beyond chance from one measurement situation to the other.

In the original statement of this theorem [4], a general practical solution was not attempted, and proof was limited to demonstrating (*a*) that proportionality of loadings cannot be attained by any orthogonal rotation whatever of two factor matrices (with the same number of factors and variables) taken at random, i.e., that it depends on the existence of real 'organic' structure in the data, and (*b*) that, if *one* position exists in which the loadings of factors in one matrix are proportional to those of corresponding factors in the other, then no other pair of orthogonal rotations exist for which this is true, i.e., there is uniqueness. It is proposed to show here how the solution may be generalized to any number of dimensions and adapted to computation. To anticipate the formal mathematical development, the problem is that of finding D, a diagonal matrix containing the several distinct proportionalities by which factors alter from experiment A to experiment B, so that we may write $V_{nA} = V_{nB}D$, where V_{nA} is the rotated position of the first factor matrix, obtained from the unrotated matrix V_{oA} by an as yet unknown transformation λ_A, and λ_A is a rectangular matrix. Thus $V_{oA}\lambda_A = V_{nA}$, and similarly $V_{oB}\lambda_B = V_{nB}$.

Before developing this, we should note that the method proposed uses the principle of parsimony involved in Thurstone's simple structure in a more comprehensive fashion. In fact, it requires the principle to hold with respect to the general field of scientific observation and inference (as, philosophically, it should) instead of merely within the universe of a single experiment or matrix. It applies it to *any* two matrices (requiring that they yield the same factors), and therefore to all. Simple structure hopes, and often finds, that what is simplest in one matrix is simplest for many; but proportional profiles *require* that all the phenomena analysed in the various matrices or experiments shall be explained by the smallest number of factors, i.e., by demonstrably similar patterns. Incidentally, the definition of

'diverse conditions' could be extended, and probably with advantage, to comparisons of *R*- and *P*-technique factorizations.

The labour of arriving at the unique proportional profiles position by simultaneous trial and error in two matrices would obviously be prohibitive; but fortunately, by the nature of the method, an analytical solution is possible, at least for errorless data. In the 1944 article [4] this was indicated for two factor problems. Before proceeding to the general solution, a digression will be made on some experimental attempts to set up the problem, since (*a*) they illustrate the theoretical relations more clearly, (*b*) they may save others from the pitfalls we encountered, and (*c*) they will indicate ways in which the theoretical solution still requires adaptation to real conditions.

III. HISTORY OF SOME EMPIRICAL EXPLORATIONS OF THE PROBLEM

The writers felt an urgent need of the method in psychological research; and the ten-year delay in further publication has been due not merely to the pressure of other work but to certain misfortunes with empirical try-outs of the method.

It has been argued [7, 9] that factor analysis in general would benefit from more *a posteriori* reasoning, as in Burt's bottle problem [2] or Thurstone's box problem [25], i.e., taking experimental situations in which the 'factors' or 'influences' are well known and seeing how various factor analytic techniques behave. In the present problem the advantages would be didactic (*a*) in demonstrating to those psychologists who are more pragmatic or philosophical than mathematical that factors are more than mathematical abstractions, (*b*) in showing whether the accidentally or experimentally produced changes in factor variance from sample to sample are large enough to produce the effects required, and (*c*) in finding whether the blurring of simple proportionality by experimental and sampling errors (in addition to the systematic distortion by selection efforts) are so large as to affect the practicability of an error-free analytical solution. The last is the most important.

Three kinds of examples, additional to the completely errorless 'pure principle' case here demonstrated, have been or are in process of trial and publication: (1) a box problem similar to Thurstone's, but with orthogonal factors and with the additional hurdle of powered relations departing from the 'pure principle' used here; (2) a similar box problem with experimental errors introduced; and (3) real and natural data, which contain both experimental and sampling error, but in which the factors are controlled and known with tolerable certainty.

The two first will be described elsewhere by Haverland [16] who has shown that the principle of proportional profiles works excellently in these cases. The third is represented by two examples which will be briefly described here and one of which is fully described elsewhere [16]. In both a search was made for a situation in which (for simplicity) *two* common factors, and two only, operated on a dozen or more variables. For the first it was decided to work on growing tomato plants, since botanists believed that two factors, representing light and nutrition levels, would be found. For help in setting up this experiment, we are greatly indebted to Dr. D. Gottlieb of the University of Illinois Horticultural Department.

Two populations, containing 200 plants in each, were grown from seed; and measurements were finally made on such variables as height of plant, thickness of stem, greenness of leaf, number of leaves, etc. In one sample the variance was made twice as great in light intensities, and in the other it was increased by 3/2 in concentration of hydroponic nutrition. The factor scores, at least as ranks, were thus known for each plant. The whole design collapsed, however, when it was found after six months that nine-tenths of the common factor variance was due to one factor only—indeed there seemed a doubt as to whether any second factor should be extracted. In this extreme factor composition no orthogonal or even oblique position of rotation could be found that would give proportionality of profiles and indeed no rotation of any kind would give a satisfactory 'light intensity' factor. Anyone repeating is advised to include more variables known botanically to be appreciably responsive to light alone.

Still seeking for living and organic (rather than merely physical) examples that would nevertheless be controllable we were encouraged by the recent discovery of drives as factors [10] to suggest that hunger and thirst be factored from the usual learning and deprivation variables on two populations of rats. This work, by Haverland [16], had larger variation of food deprivation in one sample and larger variations of water deprivation in the other. The many interesting findings cannot be described here; but it is questionable whether proportional profiles gave a meaningful solution, whereas simple structure gave easily recognizable drive patterns (best oblique, but acceptable even when orthogonal). The steps likely to produce advances beyond this situation have been indicated elsewhere [15]; but even in the purely mathematical treatment given here it would seem that a further step must eventually be taken to allow for the fact that the proportionality is only approximate. If a variable is loaded by

a factor 0·5 and the variance of that factor (still orthogonal) is doubled relative to other factors the loading will not equal 1·0. If we apply the formulae of Thomson and Ledermann [24] and Thurstone [25] for predicting the correlations to be expected after the variance of the selection variable has been changed (counting the factor as the variable on which selection is made), it will be found that not all the loadings of a factor will change in the same proportion when the factor variance is changed. However, *over small ranges* the simple proportionality is near enough, and Haverland's box example with errors [16] shows that it holds well enough to give a correct solution. Accordingly we shall here present the solution brought to that level, and leave for further invention the adjustment to non-proportionality and the other practical difficulties (e.g., obliqueness) raised by the rat problem.

IV. THE ANALYTICAL SOLUTION FOR PROPORTIONAL PROFILES, WITH ORTHOGONAL AXES

We now suppose that two experiments, A and B, with the same n variables have yielded two correlation matrices and two unrotated factor matrices. The only assumptions made are (*a*) that the matrices are of the same order, both containing n rows and k columns, and of the same rank, k, (*b*) that enough variables exist for satisfactory determination of the communalities [23], (*c*) that they occupy the same factor space, i.e., if the matrices are placed alongside to give order $n \times 2k$, the rank will remain k, and (*d*) that the variances of the true influence on these variables in the two situations are not unchanged: in fact, the variance ratios for the factors must be different from unity and from each other.

Parenthetically, the method of proportional profiles should be distinguished, in objectives and methods, from the method of aligning two factors from different studies which is implicit in the formulations of Barlow and Burt [2],[1] and from Tucker's related method [27] which has the rather different objective of synthesizing two factor studies to yield a least squares oblique solution to equation [3] below. The latter furnishes no restrictions such as are required to obtain a unique solution for the rotation; hence an additional criterion, such as simple structure or the present method (adapted to the oblique case), needs to be applied to bring the two studies marching in step to a unique halting point. A similar objective—getting maximum agreement between two studies—has been pursued by Wrigley and Neuhaus [30]: their method is distinguished from proportional profiles by being applied *after* the studies, except for the provision of common variables in the two studies. Proportional profiles on the other hand, requires the stage to be set beforehand, in order to provide a unique rotational solution dependent on controlled experimentation. It does so in order that a particular relation may be sought, which is not merely one of mathematical convenience or economy, but is posited to arise from the existence of an inherent structure.

From a rotation of a given factor matrix we obtain in the (notation above):

$$V_{oA}\lambda_A = V_{nA}, \tag{1}$$

$$\text{and} \quad V_{oB}\lambda_B = V_{nB}. \tag{2}$$

[1] *Editorial Note.* In correspondence Professor Cattell has suggested that I might indicate my own views on the problems raised. The question discussed above would seem to be a special instance of the *third* case in the paper of ours that he quotes, viz., that in which the battery of tests is the same but the sample of persons different (p. 54). We dealt with it less fully than the other two, because it had already been discussed in the references cited. For the instance he has in mind, our procedure would be to extract the initial 'bipolar' factor matrices by *weighted* summation, not by simple (i.e., not by the centroid method). With the notation there used, we should obtain

$$G_1 = F_1H_1 = L_1V_1^{\frac{1}{2}}H_1 \text{ and } G_2 = F_2H_2 = L_2V_2^{\frac{1}{2}}H_2.$$

Identity of factors is defined to mean identity of their direction-cosines (p. 52). It therefore follows that $G_2 = G_1D$. Now let $F_2 = F_1T$. Then

$$F_1T = F_2 = L_2V_2^{\frac{1}{2}} = L_1V_1^{\frac{1}{2}}H_1DH_2'.$$

Premultiplying by $V_1^{-\frac{1}{2}}L_1'$, we find $T = H_1DH_2' = V_1^{-\frac{1}{2}}L_1'L_2V_2^{\frac{1}{2}}$. The matrix, T, therefore, must then itself be factorized by weighted summation, and we at once obtain $H_1, D,$ and H_2. (This seems better than factorizing $TT' \equiv KK'$, which would give 'canonical multipliers' rather than latent roots and vectors.) The use of weighted summation avoids the difficulty mentioned by Professor Cattell, namely, that with his formula K and K^{-1}, when independently computed, are not consistent. I may add that, even with data that do not involve experimental or sampling errors, T (Professor Cattell's K) is as a rule not symmetrical, but (as noted in my earlier paper) tends rather to be triangular : cf. Burt, C., *Brit. J. Educ. Psychol.*, IX, 1939, pp. 45–71 and refs. But in general it seems desirable to determine the group factors separately and prove their similarity rather than to postulate their similarity and rotate them accordingly.—*C.B.*

The positions V_{nA} and V_{nB} are by definition such that each factor column in one matrix has loadings that are a simple multiple of those in the corresponding column of the other, assuming corresponding factors arranged in the same order. The transformation from one to the other, involved in the essential *principle of proportional profiles*, is thus simply represented by a $k \times k$ diagonal matrix D, such that:

$$V_{nA} = V_{nB}D. \tag{3}$$

Substituting from (1) and (2) into (3), we have:

$$V_{oA}\lambda_A = V_{oB}\lambda_B D, \tag{4}$$

or, keeping solely to orthogonal transformations,

$$V_{oA} = V_{oB}\lambda_B D\lambda'_A \tag{5}$$

(since $\lambda' = \lambda^{-1}$, when λ is an orthogonal matrix).

To handle the term on the right, equation (5) may more conveniently be re-written

$$V_{oA} = V_{oB}K, \tag{6}$$

$$\text{where} \quad K = \lambda_B D\lambda'_A. \tag{7}$$

Now V_{oA} and V_{oB} are the given factor matrices, defined as to number of factors, loadings, etc., by the experimental process. Consequently they provide the empirical basis necessary for calculating K and the values derived from it. Assuming K calculated,[1] let us see how to decompose it into the values required. We first multiply K by its transpose, obtaining $KK' = \lambda_B D\lambda'_A\lambda_A D\lambda'_B = \lambda_B D^2\lambda'_B$.

This expression has the typical form for the latent roots and latent vectors of a symmetrical matrix [29], D^2 being a diagonal matrix containing the latent roots and λ_B containing the latent vectors. Consequently, after extracting the latent roots and vectors in the usual way, we can obtain D, by taking the square roots of the elements in D^2. With λ_B thus determined we can obtain V_{nB}; and with D determined we can obtain V_{nA} by (3). This suffices to determine the position; but to solve completely by obtaining λ_A also it is computationally easier [2] to use

$$K'K = \lambda_A D\lambda'_B\lambda_B D\lambda_A = \lambda_A D^2\lambda'_A.$$

The solutions for λ_A and λ_B by these procedures are unique, thus confirming our earlier two-space geometrical proof of the uniqueness of the proportional profiles position [3]. It will be noted that a unique solution is not possible if our assumption is broken that the variances due to the true influences (factors) in the two situations do *not* remain unchanged. Thus if one (or more) of the diagonal values in D are unity or have the same value, one (or more) of the factors cannot be uniquely rotated.

V. A WORKED EXAMPLE

The working procedures will now be illustrated by an artificial problem involving seven variables and three factors. It has been so constructed as (1) to be free of sampling and experimental error, (2) to be susceptible of solution by the usual simple structure method, and (3) to have known proportionalities between the corresponding factors (though we shall not assume this knowledge in achieving our solution).

[1] The computation for fallible, sample-affected data has to be based on least squares. This involves minimizing the total of the squared entries in the difference matrix $V_{oA} - V_{oB}K$. This can be verified by writing these in terms of their elements, taking partial derivatives with respect to the elements of K, setting these equal to zero, and solving the resultant equations for the elements of K. This approach was suggested by Rao [20]. Turnbull and Aitken [29, p. 173] demonstrate this approximation by general methods. It has been used for related purposes by Gibson [14], Lubin [17], Mosier [18], and Tucker [27]. Gibson has suggested (personal communication) a rider whereby K can be calculated from V_{oB} and R_A only, and factored by Hotelling's iterative procedure, i.e., " by an implicit factoring which tends to select from R_A those dimensions which are represented in V_{oB}." However this leaves the communality problem out of consideration. Unfortunately, for fallible experimental data the solutions for K are not symmetrical, and $V_{oA} = V_{oB}K$ yields a different result from $V_{oB} = V_{oA}K^{-1}$. K^{-1} may be obtained from the expression $(V_{oA'}V_{oA})^{-1}V_{oA'}V_{oB}$. The two K values can be computed as estimates and KK' calculated. This will not be an exactly symmetrical matrix; but can be made so by averaging the corresponding elements on opposite sides of the diagonal. Tucker's device, using mutual regression [27], is perhaps somewhat neater. Nil results may be checked by $\lambda_B = K\lambda_A D^{-1}$.

[2] Another solution for λ_A follows from [7], which gives: $\lambda_A = K'\lambda_B D^{-1}$.

Relatively small differences of variance, such as might exist in practical examples, were chosen, such that the factor loadings in the second matrix are respectively 2/3rds, 9/8ths, and 6/5ths of those for corresponding variables in the first. To give a local habitation and a name to the problem, we can imagine that the correlation matrix represents relations between seven tests involving the manipulation of mechanical puzzles, and that success is determined by factors of mechanical aptitude (F1), reasoning ability (F2), and manual dexterity (F3). The two matrices may represent data from boys and girls respectively, F1 having less variance in the boys and F2 and 3 in the girls. Let us begin, as it were, from behind the scenes with the two final, rotated matrices, representing the simple structure and the functionally meaningful position just described. These are set out in Table I.

TABLE I. FACTOR MATRICES

(As known to be structured)

A. Boys					B. Girls				
Tests	F_{1nA}	F_{2nA}	F_{3nA}	$h^2{}_A$	Tests	F_{1nB}	F_{2nB}	F_{3nB}	$h^2{}_B$
1	·60	−·40	−·10	·53	1	·40	−·45	−·12	·37
2	−·50	−·80	·10	·90	2	−·33	−·90	·12	·93
3	−·10	·80	·00	·65	3	−·17	·90	·00	·84
4	·00	·50	−·60	·61	4	·00	·56	−·72	·83
5	·90	−·10	·00	·82	5	·60	−·11	·00	·37
6	−·60	·00	·70	·85	6	−·40	·00	·84	·87
7	·05	·00	·80	·64	7	·03	·00	·92	·92

For uniformity we have labelled the factors the F_{nA}'s and the F_{nB}'s. It will be observed that the required proportionalities exist between the loading in each pair of columns, and that each factor in each matrix has a position of tolerable good simple structure, having three or more variables in the hyperplane for each factor (generally within $\pm 0{\cdot}12$ but in one instance $0{\cdot}17$); and each variable with at least one hyperplane loading [25].

The correlation matrices reconstructed from these two factor matrices, by taking the inner products of the rows ($V_{nA}V_{nA'}$), are shown in Table II.

TABLE II. OBTAINED CORRELATION MATRICES

R_A (Boys)								R_B (Girls)							
Tests	1	2	3	4	5	6	7	Tests	1	2	3	4	5	6	7
1	—							1	—						
2	·01	—						2	·26	—					
3	−·38	−·59	—					3	−·47	−·75	—				
4	−·14	−·46	·40	—				4	−·17	−·59	·50	—			
5	·58	−·37	−·17	−·05	—			5	·29	−·10	−·20	−·06	—		
6	−·43	·37	·06	−·42	−·53	—		6	−·26	·23	·07	−·60	−·24	—	
7	−·05	·06	−·01	−·48	·04	·53	—	7	−·10	·11	−·01	−·69	·02	·79	—

Taking these correlation matrices as our starting point, let us try to arrive at a meaningful rotated solution (*a*) by the new principle of proportional profiles and (*b*) by rotation for simple structure. A centroid factorization of the correlation matrices yields the following unrotated factor matrices (Table III).

TABLE III. UNROTATED FACTOR MATRICES

V_{oA} (Boys)					V_{oB} (Girls)				
Tests	F_{oA_1}	F_{oA_2}	F_{oA_3}	h^2	Tests	F_{oB_1}	F_{oB_2}	F_{oB_3}	h^2
1	34	−66	01	55	1	31	51	18	39
2	59	29	−68	89	2	71	32	−58	94
3	−63	25	45	66	3	−69	−60	14	86
4	−75	−19	−07	60	4	−88	21	05	82
5	22	−77	45	84	5	15	29	51	37
6	34	86	11	87	6	49	−78	−19	88
7	50	37	51	65	7	63	−70	22	94

Let us first solve from this point by the proportional profiles method From (6) we know

$$V_{oA} = V_{oB}K,$$

$$\text{whence} \qquad K = V_{oB}^{-1}V_{oA}. \tag{8}$$

Our next step is to get K by equation (8). However, since V_{oB} is not a square matrix, it cannot have an inverse.[1] But a device is possible which we may call the symmetrizing method. This has the advantage that it is certain to yield D; whereas the trimming method (see footnote 2) sometimes fails. By this approach we aim first at a symmetric matrix S, such that $S = V'_{oB}V_{oB}$. In this calculation the *whole* V'_{oB} and V_{oB} is used, yielding, in this case, a 3 × 3 matrix. Note that, since $S^{-1} = V_{oB}^{-1}(V'_{oB})^{-1}$, and $K = V_{oB}^{-1}V_{oA} = V_{oB}^{-1}(V'_{oB})^{-1}V_{oB}V_{oA}$, therefore $K = S^{-1}V'_{oB}V_{oA}$.

This use of the square non-singular matrix S thus permits us to solve[2] for K; and we then obtain:

$$S = \begin{bmatrix} \cdot4342 & \cdot0396 & \cdot2225 \\ \cdot0396 & \cdot5169 & \cdot0379 \\ \cdot2225 & \cdot0379 & 1\cdot4743 \end{bmatrix} \quad K = \begin{bmatrix} \cdot8638 & \cdot0070 & \cdot0162 \\ \cdot0854 & -\cdot8293 & -\cdot3971 \\ \cdot1035 & -1\cdot0279 & 1\cdot0563 \end{bmatrix} \quad KK' = \begin{bmatrix} \cdot7465 & \cdot0615 & \cdot0993 \\ \cdot0615 & \cdot8527 & \cdot4418 \\ \cdot0993 & \cdot4418 & 2\cdot1831 \end{bmatrix}$$

The latent roots and vectors were obtained on the electronic computor by the method of Truman Kelley, adapted and reported by Wrigley and Neuhaus [30]:

$$D^2 = \begin{bmatrix} 2\cdot33 & & \\ & \cdot76 & \\ & & \cdot70 \end{bmatrix} \quad D = \begin{bmatrix} 1\cdot53 & & \\ & \cdot87 & \\ & & \cdot84 \end{bmatrix} \quad \lambda_B = \begin{bmatrix} \cdot0714 & \cdot8048 & \cdot5893 \\ \cdot2895 & \cdot5487 & -\cdot7843 \\ \cdot9545 & -\cdot2266 & \cdot1938 \end{bmatrix}$$

[1] Equation (8) and those at the end of the paragraph cannot validly be written because they involve the imaginary concept of the inverse of a rectangular matrix. Since we eventually advocate a different approach, it may be asked why this initial [illegible] with its inadequate 'matrix trimming' device, is retained. It is retained because it is logically more direct and easier for non-mathematicians to follow. For, although a rectangular V_{oB}^{-1} cannot be used, a solution could evidently be obtained by trimming it to a square matrix, since the transformation applied to k variables (rows) is applied equally to all; at least this would be so with non-fallible data, though with experimental and sampling errors no such sample of rows could be truly representative of the whole matrix. There the only satisfactory solution is the least squares fit for the whole matrix (see footnote 1). Even with exact data ambiguities arise from this squaring of the rectangle or 'trimming', as will be seen below. But for maximum illumination it has seemed best to carry through both the cruder 'trimming' and the neater 'symmetrizing' procedures for comparison, though the former will be carried out only in footnotes, since it is not advocated in practice.

By the direct 'trimming' method we first trim V_{oB} and V_{oA} to three rows each and then take the inverse of the first (see (6)) as follows:

$$V_{oB}^{-1} = \begin{bmatrix} -13\cdot49 & -7\cdot98 & -15\cdot72 \\ 13\cdot38 & 7\cdot46 & 13\cdot68 \\ -9\cdot13 & -7\cdot38 & -11\cdot70 \end{bmatrix}$$

which, post-multiplied by the square matrix of V_{oA} (trimmed), becomes:

$$K = \begin{bmatrix} \cdot609 & 2\cdot657 & -1\cdot782 \\ \cdot328 & -3\cdot248 & 1\cdot221 \\ -\cdot090 & \cdot960 & -\cdot335 \end{bmatrix}$$

[2] As will be seen, this result differs decidedly from the value for K found from the 'trimmed' matrices.

The values for V_{nA} and V_{nB} are given by eqns. (2) and (3); and the results are shown in Table IV. λ_A can be reached, as indicated above, either by calculating $K'K$ and proceeding therefrom as from KK' (which will also give a check on D^2) or by eqn. (1), p. 86.

TABLE IV. PROPORTIONAL PROFILES SOLUTION
(Symmetrizing Approach)

$V_{nA} = V_{nB}D$ (Boys)				$V_{nB} = V_{oB}\lambda_B$ (Girls)			
Tests	F_{nA_1}	F_{nA_2}	F_{nA_3}	Tests	F_{nB_1}	F_{nB_2}	F_{nB_3}
1	·53	−·42	−·15	1	·34	−·48	−·18
2	−·63	−·77	·05	2	−·41	−·88	·06
3	−·14	·80	·08	3	−·09	·92	·09
4	·08	·52	−·56	4	·05	·60	−·67
5	·89	−·14	−·03	5	·58	−·16	−·04
6	−·57	·01	·72	6	−·37	·01	·86
7	·08	·07	·80	7	·05	·07	·96

Except for errors from rounding, these values are the same as in Table I. Even in F^1 and F_2, where the approximation is rougher, one has no difficulty in recognizing the similar patterns. A further example since worked by Haverland with improved mechanics of computing furnished the original values correct to four decimal places [16]. Furthermore, the values for D—1·53, 0·87, and 0·84—are close enough for all practical purposes [1] to the original, namely, 1·50, 0·89, and 0·83; and the closeness of these ratios to unity has provided a severer test of the method than we hope will normally be demanded.

TABLE V. PROPORTIONAL PROFILES SOLUTION
(' Trimming Approach ')

V_{nA} (Boys)				V_{nB} (Girls)			
Tests	F_{nA_1}	F_{nA_2}	F_{nA_3}	Tests	F_{nB_1}	F_{nB_2}	F_{nB_3}
1	·58	−·46	·01	1	·32	−·52	−·12
2	−·58	−·74	−·01	2	−·46	−·85	·12
3	−·00	·81	·00	3	−·05	·92	·00
4	·16	·48	−·59	4	·10	·53	−·73
5	·88	−·15	·20	5	·57	−·20	·00
6	−·73	·06	·57	6	−·40	·09	·85
7	−·12	·01	·80	7	·01	·02	·97

The agreement is almost as good as by the other approach; but the D matrix failed, and the V_{nA} matrix had to be calculated through $K'K$. The D matrix gave, not the known ratios, but 0·89, 4·79 and 0·00! presumably taking only three rows of a matrix is likely to reduce its rank, since four of the seven variables must be linear combinations of the others; and we are unlikely to hit by chance on those that would preserve the rank. Accordingly, since ' symmetrizing ' gives more immediately what is required and avoids the need to find a pseudo-inverse, it seems the better procedure.

Now let us compare with the above proportional profiles determinations the simple structure position as reached by the usual trial and error search. This is done primarily to

[1] It is interesting to compare this result with that found by the ' trimmed ' matrix K. The V_{nA} and V_{nB} values so found are:

compare the relative computational times required and the role of ambiguity and approximation in the two methods. A blind rotation for simple structure was carried out from Table III by an experienced worker. Rotations proceeded steadily to a satisfactory structure, obtained after six over-all rotations for V_{nA}, and five for V_{nB} (Table VI).

TABLE VI. SIMPLE STRUCTURE

V_{nA} (Boys)				V_{nB} (Girls)			
Tests	F_{nA_1}	F_{nA_2}	F_{nA_3}	Tests	F_{nB_1}	F_{nB_2}	F_{nB_3}
1	·65	−·35	−·06	1	·40	−·47	−·08
2	−·14	·80	−·04	2	−·36	−·90	·07
3	−·45	−·83	·11	3	−·16	·90	·00
4	·00	·45	−·63	4	·11	·56	−·71
5	·92	·01	·08	5	·59	−·11	·06
6	−·65	−·01	·67	6	−·50	·02	·79
7	−·01	·06	·81	7	−·09	·00	·96

Blind rotation has thus succeeded in finding the original simple structure, though others *might* find an alternative structure. In their agreement with the true position (Table I) the simple structure (Table V) and the proportional profiles methods (Tables IV and V) are equally good. But in the wider applicability and in the analytical solution the proportional profiles solution can claim superiority, at least when restriction to the orthogonal case is overcome.

VI. SUMMARY

1. The assumption is made that, when (and only when) factors correspond to organic influences inherent in the data, will their loadings alter as a whole, by ratios peculiar to each, from one situation to another in which the factor as a whole is differently involved.

2. It has been shown in the orthogonal case that, when this relation exists, it can be found only by one unique paired rotation of the two factor matrices, and that this position can be discovered with certainty by an analytic solution.

3. An example has been worked on errorless data, in which the solution was made to be simultaneously the simple structure and the proportional profiles position. Re-working from the correlation matrix, this position was correctly found, without ambiguity, by two methods. The former required about seven hours of computation and drawing ; the latter about two hours of tape-punching and ten minutes of electronic computing.

4. Work in progress confirms that the analytic solution can be obtained with great accuracy for errorless data, and also, with acceptable accuracy, for a box problem with introduced error. However, an actual psychological experiment [16], in which control was attempted of the factors believed to account for the variance, gave poor agreement of proportional profiles with simple structure and the presumed psychological structure.

5. Even without this example, it would be evident that the method awaits the following improvements for its effective use. (i) Adaptation to oblique axes; additional restrictions will be necessary to obtain a unique solution in this case. (ii) Allowance for the fact that, as implied in the work of Thomson and Ledermann [24], variance change will not produce a simple proportional change in the loadings of all variables, but a change which is partly a function of the individual variable loading. (iii) Allowance in individual loadings for the fact that the experimenter will often be compelled to produce his experimental change of factor variance by selecting on a variable highly loaded in the factor, instead of on the factor itself.

REFERENCES

1. Bargmann, R. (1954). 'Signifikanzuntersuchungen der Einfachen Struktur in der Faktoren-Analyse.' Sonderdruck. Physica-Verlag, Würzburg.
2. Barlow, J. A. and Burt, C. (1954). 'The identification of factors from different experiments.' *Brit. J. Stat. Psychol.*, VII, 52-53.
3. Carroll, J. B. (1953). 'An analytical solution for attaining simple structure in factor analysis.' *Psychometrika*, XVIII, 23-28.
4. Cattell, R. B. (1944). 'Parallel proportional profiles and other principles for determining the choice of factors by rotation.' *Psychometrika*, IX, 267-283.
5. Cattell, R. B. (1946). 'Simple structure in relation to some alternative factorizations of the personality sphere.' *J. Gen. Psychol.*, XXXV, 225-238.
6. Cattell, R. B. (1946). *The Description and Measurement of Personality.* New York : World Book Co.
7. Cattell, R. B. (1952). *Factor Analysis : An Introduction and Manual for the Psychologist and Social Scientist.* New York : Harper.
8. Cattell, R. B. (1955). 'The principal replicated factors discovered in objective personality tests.' *J. Abnorm. Soc. Psychol.*, L, 291-314.
9. Cattell, R. B. (1955). 'Growing points in factor analysis.' *Austral. J. Psychol.*, VI, 105-140.
10. Cattell, R. B. and Miller, A. (1952). 'A confirmation of the ergic and self-sentiment structures among dynamic traits attitude variables by R-technique.' *Brit. J. Psychol.*, XLI, 280-294.
11. Cattell, R. B. and Gruen, W. (1955). 'Measurement of the primary personality factors in children by means of objective tests.' *J. of Person.*, XXIII, 460-478.
12. Eysenck, H. J. (1950). 'Criterion rotation, an application of the hypothetico-deductive method to factor analysis.' *Psychol. Rev.*, LVII, 38-65.
13. French, J. W. (1953). 'The description of personality measurements in terms of rotated factors.' Princeton: Educational Testing Service.
14. Gibson, W. A. (1953). 'A least squares solution for case IV of the law of comparative judgment.' *Psychometrika*, XVIII, 15-21.
15. Goodman, C. H. (1943). 'A factor analysis of Thurstone's sixteen primary mental abilities tests.' *Psychometrika*, VIII, 141-151.
16. Haverland, E. M. (1954). 'The application of an analytical solution for proportional profiles rotation to a box problem and to the drive structure in rats.' Ph.D. Thesis, University of Illinois Library, Urbana, Ill.
17. Lubin, A. (1950). 'A note on " criterion analysis ".' *Psychol. Rev.*, LVII, 54-57.
18. Mosier, C. I. (1939). 'Determining a simple structure when loadings for certain tests are known.' *Psychometrika*, IV, 149-162.
19. Neuhaus, J. O., and Wrigley, C. (1954). 'The quartimax method : an analytical approach to orthogonal simple structure.' *Brit. J. Statist. Psychol.*, VII, 88-92.
20. Rao, C. R. (1952). 'Advanced statistical methods in biometric research.' New York: Wiley.
21. Sandler, J. (1952). 'A technique for facilitating the rotation of factor axes, based on equivalence between persons and tests.' *Psychometrika*, XVII, 223-229.
22. Saunders, D. R. (1953). 'An analytic method for rotation to orthogonal simple structure.' Princeton : Educational Testing Service Rept.
23. Thomson, G. H. (1952). 'The Factorial Analysis of Human Ability.' London: Univ. London Press. 5th edn.
24. Thomson G. H., and Ledermann, W. (1939). 'The influence of multivariate selection on the factorial analysis of ability.' *Brit. J. Psychol., Gen. Sect.*, XXIX, 288-306.
25. Thurstone, L. L. (1947). 'Multiple Factor Analysis.' Chicago: Univ. Chicago Press.
26. Thurstone, L. L. (1953). 'Analytical method for simple structure.' Adv. Publ., No 6, Psychom. Lab., Chapel Hill, N.C.
27. Tucker, L. R. (1951). 'A method of synthesis for factor analysis studies.' Washington: Dept. of the Army, Adjutant General's office, P.R.S. Rept. No. 984.
28. Tucker, L. R. (1953). 'The objective definition of simple structure in linear factor analysis.' *Psychometrika*, XX, 209-226.
29. Turnbull, H. W., and Aitken, A. C. (1932). 'An introduction to the Theory of Canonical Matrices.' London : Blackie & Son.
30. Wrigley, C. F., and Neuhaus, J. O. 'The matching of two sets of factors.' (In press.)
31. Wrigley, C. F., and Neuhaus, J. O. 'The use of an electronic computer in principal axes factor analyses.' *J. Educ. Psychol.* (In press.)

Vol. VIII
Part II

The British Journal of Statistical Psychology

November
1955

MECHANISMS INVOLVED IN GROUP PRESSURES ON DEVIATE-MEMBERS

By HERBERT A. SIMON and HAROLD GUETZKOW
Carnegie Institute of Technology

I. *Problem.* II. *Variables and Equations of the Deviate-Member Model.* III. *Comparison of Deviate-Member and Aggregate Models.* IV. *Cohesiveness and Rejection.* V. *Summary.*

I. PROBLEM

Leon Festinger and his associates have stated and tested a number of propositions about communication processes in small groups [1]. The purpose of this paper is to carry a step further the synthesis of a segment of these propositions into an interrelated system. First, we shall translate some of Festinger's propositions into a more rigorous form, and examine the correspondence between our equations and his verbal statements. Secondly, we shall examine an empirical study by Schachter [2] in its bearing upon our model.

For many purposes it is convenient to describe a group as though it were an aggregate which may be characterized by merely averaging or summing the individual characteristics of all its members. But as we attempt to explore the internal processes within groups, we find it useful to differentiate one member of the group (e.g., the leader) from the others.[1] Many of the propositions Festinger has formulated for group pressures belong to the first type; they are concerned entirely with aspects of group behaviour that can be treated as averages or aggregates for the group as a whole. Elsewhere [3] we have examined this part of Festinger's theoretical scheme. The present paper will undertake a parallel analysis of the remainder of Festinger's system—those propositions that involve specific reference to a particular member, A, of the group. The member selected for attention is a *deviate-member;* and the principal variables in the system represent the attitudes and responses toward the deviate-member of the remaining members. In a later section we shall comment on the relation between the two parts of Festinger's system, and between the models we have constructed to represent them.

Festinger states five hypotheses about communication resulting from pressures toward uniformity that describe the behaviour of deviate-members. Of these two contain references to the communication among individuals in a group; one is concerned with other potential membership groups; the remainder refer to changes in composition of the group, e.g., through pushing members out of it. Our procedure will be to set forth our system, and then to relate the model to Festinger's verbal hypotheses.

II. VARIABLES AND EQUATIONS OF THE DEVIATE-MEMBER MODEL

The five deviate-member hypotheses of Festinger—those numbered 2a, 2b, 2c, 4a, and 4c—are stated in terms of eight variables in addition to time. Consider a group of n members and a particular member, A. The variables refer to averages over the group *excluding* A.

R: The *relevance* to the group of uniformity in respect of the topic under discussion.

$P_A(t)$: The average *Pressure* upon members of the group (excluding A) to communicate *to* A at time t.

[1] The importance of the distinction has been recognized by the economist in applying index numbers in his analysis of economic behaviours. We speak of an 'index of wholesale prices' only if we are interested just in the general movement of prices, and if all prices tend to move together. If either of these conditions fails, the index becomes correspondingly meaningless. Thus, it is a pragmatic question as to when we can employ an aggregate model, or when we must formulate a more complete model in which the definition of the variables involves less aggregation.

$D_A(t)$: The *Discrepancy* between the opinion *of A* and the average opinion of the rest of the group, *as perceived by the rest at time t.*

$C_A(t)$: The *cohesiveness* of the group *with A*—that is, the average intensity of the desire to retain A as a member of the group at time t.

$L_A(t)$: The *perceived* receptivity (' listening ') of A to influence—that is, the average estimate by the other members of his willingness to be influenced by them at time t.

$U_A(t)$: The average pressure toward uniformity of the group with A, *as felt by the group* at time t.

$C(t)$: The *cohesiveness* of the group, exclusive of the deviate member—that is, the average intensity of the desire of the other members to remain in the group.

We now hypothesize the following system of equations:

$$P_A(t) = P_A[D_A(t),\ U_A(t)];\ \text{(see footnote (1))}. \tag{1.1}$$

$$\frac{dL_A(t)}{dt} = \phi[D_A(t),\ L_A(t)]. \tag{1.2}$$

$$\frac{dC_A(t)}{dt} = g_A[D_A(t),\ U_A(t),\ C_A(t),\ L_A(t)]. \tag{1.3}$$

$$U_A(t) = U_A[C_A(t),\ R,\ C]. \tag{1.4}$$

Since Festinger has defined cohesiveness as ' the resultant of all the forces acting on members to remain in the group ', the level of $C(t)$ at which members are indifferent as to the choice between remaining in or leaving the group could be taken as a natural zero for this cohesiveness variable. Analogously, $C_A(t)$ may be defined as the average *net* attitude of the group (other than A) towards A's remaining in the group. Then positive C_A would mean a net preference for his remaining, negative C_A would mean *rejection*—a net preference for his leaving.

The four equations constituting our model postulate that the changes in P_A and U_A are instantaneous, while the changes in L_A and C_A take place gradually through time, as functions of the variables appearing on the right-hand sides of their respective equations. A discussion of the nature of such systems of algebraic and differential equations is presented in the appendix.

We are now ready to examine Festinger's five hypotheses concerned with the deviate member:

Hypothesis 2*a:* The pressure to communicate about ' item x ' to a particular member of the group will increase as the discrepancy in opinion between that member and the communicator increases.

This hypothesis states a relationship between D_A and P_A, as defined in our equation (1.1) and asserts that $\dfrac{\partial P_A}{\partial D_A} > 0$.

Hypothesis 2*b:* The force to communicate about ' item x ' to a particular person will decrease to the extent that he is perceived as not a member of the group or to the extent that he is not wanted as a member of the group.

This hypothesis states a relationship between P_A and C_A, via U_A (eqns. (1.1) and (1.4)) and asserts that $\dfrac{\partial P_A}{\partial U_A} > 0$, $\dfrac{\partial U_A}{\partial C_A} > 0$.

Hypothesis 2*c:* the force to communicate ' item x ' to a particular member will increase the more it is perceived that the communication will change that member's opinion in the desired direction.

The present system of equations does not translate the full content of this proposition, in particular that part of it which refers to the ' desired direction ' of change in opinion. A closely related proposition stated by Schachter [2] plays an important role in his analysis. Schachter makes the assumption ' that perceived difference increases with discussion '

[1] Read: The magnitude of the pressure upon group members to communicate to A at time t is a function of the discrepancy of A's opinion from the rest of the group, and of the pressure toward uniformity.

when the actual discrepancy remains constant [2, p. 201]. In the present system of equations, the mechanism that is postulated to produce this kind of behaviour is slightly different. We assume that, as discussion proceeds, the perceived receptivity of A to influence (L_A) declines if the perceived discrepancy of opinion persists. This is stated in equation (1.2) with $\frac{\partial \phi}{\partial D_A} < 0$, $\frac{\partial \phi}{\partial L_A} < 0$. Further, we assume that the lower is L_A, the lower is the equilibrium value of C_A. This is expressed by equation (3.3), with $\frac{\partial g_A}{\partial C_A} < 0$, $\frac{\partial g_A}{\partial L_A} > 0$.

By means of these additional assumptions, a high persistent value of D_A depresses L_A (eqn. 1.2); this reduces C_A (eqn. 1.3), with a consequent reduction in U_A (eqn. 1.4), and simultaneously in P_A (eqn. 1.1). *Thus, if the perceived receptivity of A to influence declines, so also will the pressure to communicate to him.*

Hypothesis 4a: The tendency to change the composition of the psychological group (pushing members out of the group) increases as the perceived discrepancy in opinion increases.

This hypothesis relates C_A to D_A (eqn. 1.3). It asserts that $\frac{\partial g_A}{\partial D_A} < 0$.

Hypothesis 4b: When non-conformity exists, the tendency to change the composition of the psychological group increases as the cohesiveness of the group increases and as the relevance of the issue to the group increases.

This relates C_A to R and C via U_A and asserts that $\frac{\partial g_A}{\partial U_A} < 0$; $\frac{\partial U_A}{\partial R} > 0$; $\frac{\partial U_A}{\partial C} > 0$ (eqns. (1.3) and (1.4)).

Our system now consists of the equations (1.1) to (1.4) together with certain assumptions about the signs of partial derivatives of the functions that appear in them. For convenience, we collect the latter assumptions.

$P_D > 0$	(1.1 a)			$P_U > 0$	(1.1 b)		
$\phi_P < 0$	(1.2 a)			$\phi_L < 0$	(1.2 b)		
$g_D < 0$	(1.3 a)	$g_U < 0$	(1.3 b)	$g_C < 0$	(1.3 c)	$g_L > 0$	(1.3 d)
$U_C > 0$	(1.4 a)	$U_{R'} > 0$	(1.4 b)	$U_{C'} > 0$	(1.4 c)		

For notational simplicity, we have suppressed the subscript A, and employed subscripts to denote partial differentiation, designating the two variables without subscripts, R and C, by R' and C' respectively.

Deviation and Rejection: The Schachter Experiment. A group of persons, including several confederates or ' paid participants ', is brought together for discussion. The variable D_A, where A is one of the confederates (the ' deviate '), is determined by the (pre-arranged) behaviour of the confederate, and hence remains constant. Different initial values of R and C are induced by the instructions given at the outset of the experiment. The cohesiveness of the groups was checked at the end of the experiment and was found constant (see (2) Table II, p. 194); R is assumed to remain constant for the forty-five minutes of the discussion. Direct measurements of C and C_A were also made at the end of the trial. During its course, measurements were made of $P_A(t)$. These were average values over ten minute intervals.

The Schachter experiment provides us with a test of the postulated model in the following respects:

(1) We can test whether differences in the parameter values, C and R, and D_A produced the predicted differences in the terminal values of C_A.

(2) We can test whether the time paths of $P_A(t)$ are consistent with those predicted by the model.[1]

Eliminating U_A by equation (1.4) and regarding D_A, R, and C as parameters, we obtain the equations:

$$\frac{dL_A}{dt} = \phi(D_A, L_A). \tag{2.1}$$

$$\frac{dC_A}{dt} = g_A[D_A, U_A(C_A, R, C), C_A, L_A]. \tag{2.2}$$

[1] Schachter, in his paper, attempts to derive inferences about the time paths of the variables from assumptions about the signs of certain partial derivatives. Because the basic equations of his model are algebraic rather than differential equations, he can do this only by introducing the *ad hoc* assumption ' that *perceived* difference increases with discussion time '. It is difficult to know precisely what assumptions underlie the time paths displayed in his Fig. 3. By mathematizing the assumption, as was done following Hypothesis 2c, it is possible to know exactly what assumptions are made.

Consider now the direction field of the system in the (C_A, L_A) plane, with the parameters held constant.[1] Since C_A does not appear in (3.5) we have:

$$\left[\frac{\delta L_A}{\delta C_A}\right]_{\phi=0} = 0. \tag{2.3}$$

$$\left[\frac{\delta C_A}{\delta L_A}\right]_{g_A=0} = -\frac{g_L}{g_U U_C + g_C} > 0. \tag{2.4}$$

There are grounds for a plausible assumption which enables us to say that these curves have a particular shape—namely, that $\frac{\delta C_A}{\delta L_A}$ (eqn. 2.4) approaches zero for very large and very small values of C_A. Then the direction field will have the general form depicted in Fig. 1. The argument for this is as follows: Equation (2.2) describes the change in the cohesiveness of the group with A. It seems reasonable that this mechanism is subject to saturation, i.e., that when C_A reaches very high levels, it will not be driven much higher by further decreases in D_A. Likewise, when C_A reaches very low levels, it will not be driven much lower by further increases in D_A. Typical paths for the system are shown from initial positions A, B, C, and D. There is a single stable equilibrium for the system at E.

Now, referring to equation (2.2), and remembering our assumptions as to signs, we find that an increase in any one or more of the parameters D_A, R, and C corresponds to a shift to the left of $g_A = 0$; while a decrease produces a shift to the right of $g_A = 0$. Similarly, from (2.1) an increase in D_A means a downward shift of $\phi = 0$ and a decrease in D_A an upward shift. It follows that an increase in one or more of the parameters shifts E to the left (and downwards if D_A increases); while a decrease in the parameters shifts E to the right (and upward). Fig. 2 shows the equilibrium curves for 'low' (solid line) and 'high' (broken line) values of the parameters, respectively.

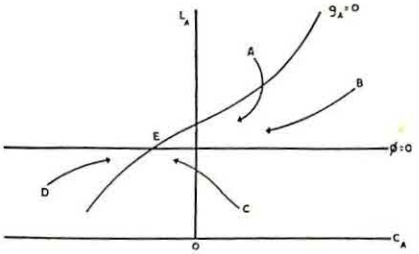


FIG. 1.—Possible Paths of the Systems
Arrows show possible paths (i.e., simultaneous changes in L_A and C_A) for various initial points, A, B, C and D. The system will be in equilibrium when it reaches E; for then, $\phi = 0$ (see equation 2.1) and $g_A = 0$ (equation 2.2); hence, L_A and C_A will remain constant.

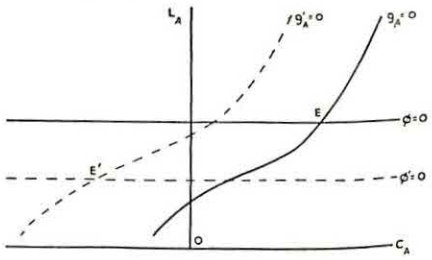


FIG. 2.—Positions of Equilibrium
E and E′ are equilibrium positions of the system corresponding to different values of the parameters appearing in equations (2.1) and (2.2).

For the particular values selected here, it may be seen that the equilibrium value of C_A is positive with low parameter values, but negative with high: that is, a confederate will be rejected by the group $[C_A(T) < 0]$ if he persists in a very deviant opinion (D_A large), if the group is highly cohesive (C large) and if the item under discussion is highly relevant to the group's goal (R large). Under the opposite conditions, the confederate will not be rejected. These conclusions are all consistent with the data of Schachter's experiment [2, pp. 198 and 199, Tables III and VI].[2]

It will simplify matters if we regard the system as approximately linear, and measure the variables as deviations from the equilibrium values. That is, we treat the partial derivatives as constant, and take zero as the equilibrium value for all the variables. The system then becomes:

$$P_A = P_D D_A + P_U U_A. \tag{2.5}$$

$$\frac{dL_A}{dt} = \phi_D D_A + \phi_L L_A. \tag{2.6}$$

$$\frac{dC_A}{dt} = g_D D_A + g_U U_A + g_C C_A + g_L L_A. \tag{2.7}$$

$$U_A = U_C C_A + U_{R'} R + U_{C'} C. \tag{2.8}$$

[1] For the methods employed here, see Lester R. Ford, *Differential Equations*, pp. 9-12, and A. A. Andronow and C. E. Chaikin, *Theory of Oscillations*, pp. 8-12, 182-193, and 203-208.

[2] It should be noted that the measures of C_A in Schachter's experiment did not determine the zero point (i.e., the line between acceptance of A and rejection). The data show that the direction of shift of E is that predicted by the model.

Four possible paths of the system, originating at points *a*, *b*, *c*, *d*, respectively, are shown in Fig. 3. Because we have taken the equilibrium point of the system as origin, a shift to the right of the initial point of the system is precisely equivalent to a shift to the left of $g_A = 0$ in the previous representation of the model. Hence *a*, *b*, and *c* may be regarded as three possible initial points, with the same initial values of L_A, but with progressively larger values of C, and/or R: (these larger values, as we have seen, shift g_A to the left relative to a fixed origin, or the initial point to the right relative to an origin at E).

From the shapes of the paths originating at *a*, *c*, and *b*, we conclude:

Case (1): If C and R are low enough, C_A will increase during the trial from its initial value towards the equilibrium value (Path *a*);

Case (2): If C and R are high enough, C_A will increase during the trial from its initial value toward the equilibrium value (Path *c*) ;

Case (3): For some intermediate range of values of C and R, C_A will first increase and then decline during the trial (Path *b*).

Now for a constant value of D_A, we see from (2.5) that the changes in P_A will be proportional to the changes in U_A, and from (2.8) that for constant R and C, the changes in U_A will be proportional to the changes in C_A. We conclude that the time shape of the curve $P_A(t)$ will be the same as that shown in Fig. 3 for $C_A(t)$.

It is in this way that we would interpret the data presented in Schachter's Table VII [2, p. 203] which measures P_A by the mean number of communications addressed to the ' deviate ' during each of four ten-minute periods in the course of the experiment. For higher values of R and C in the Schachter experiment, $P_A(t)$ fell in Case 3, above. When C or R were at lower values, $P_A(t)$ had a time path corresponding to Case 1. Case 2 was not observed, suggesting that despite Schachter's laudable attempt to " reproduce the variables and phenomena under study with greater intensity in a purportedly ' real-life situation ' " [2, p. 195], he had not succeeded in reaching high enough levels of C and R to produce the second case. However, he was successful in producing interesting qualitative differences among the time paths of $P_A(t)$ by using his four experimental combinations of high and low levels of C and R.

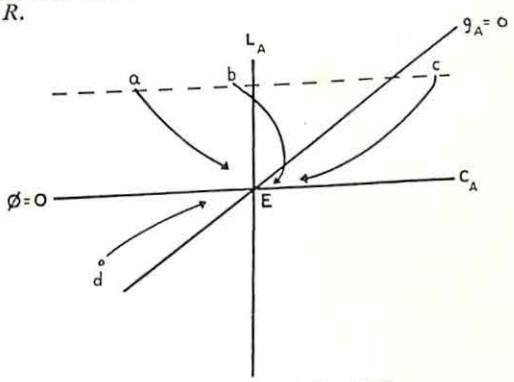


FIG. 3.—Shape of Time Path.
The arrows illustrate dependence of the shape of the time path upon the values of C and R.

III. COMPARISON OF DEVIATE-MEMBER AND AGGREGATE MODELS

As mentioned in an earlier paper [3], we have developed a model encompassing five propositions stated by Festinger [1] which deal with aggregate relations, i.e., are concerned with the members of a group taken as a whole. The model consists of the following mechanisms:

$$\frac{dD}{dt} = f[P(t), L(t), D(t)], \tag{3.1}$$

$$P(t) = P[D(t), U(t)], \tag{3.2}$$

$$L(t) = L[U(t)], \tag{3.3}$$

$$\frac{dC}{dt} = g[D(t), U(t), C(t)], \tag{3.4}$$

$$U(t) = U[C(t), R], \tag{3.5}$$

where the variables are defined as follows:

$D(t)$: The perceived *discrepancy* of opinion on an issue among members of a group at time t;

$P(t)$: *Pressure* upon members of the group to communicate with each other at time t;

$L(t)$: Receptivity ('*Listening*') of members of the group to influence by communications from other members at time t;

$C(t)$: *Cohesiveness* (identical with the corresponding variable in the deviate-member model).

$U(t)$: Pressure felt by the group to achieve *uniformity* of opinion, i.e. to reduce perceived discrepancy of opinion at time t;

R: *Relevance* (identical with the corresponding parameter in the deviate-member model).

Now let us contrast this aggregate model with the deviate model. The variables in the two models are quite analogous; but the equations are somewhat different.

It will be noticed that the mechanism postulated in equation (1.1) is the same as in (3.2), but in the former case aggregated only over the subgroup excluding A. The mechanism in (1.3) is similar to that postulated in (3.4), but has been elaborated slightly to incorporate the effect of L_A—perceived receptivity of A to influence—upon C_A—the cohesiveness of the group with A. This mechanism would not produce essential differences in the behaviour of the system unless, as in the Schachter situation, there were forces tending toward subgroup formation.

The mechanism that makes the presence of L_A in (1.3) important is equation (1.2). This states that, when there is a discrepancy between A's opinion and that of other members,

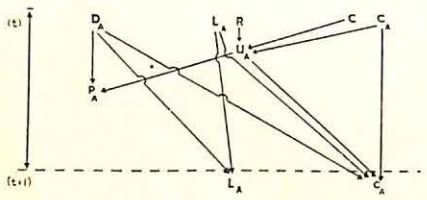


FIG. 4a.—Deviate-Member Model

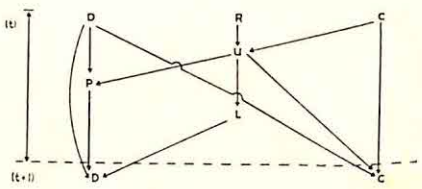


FIG. 4b—Aggregate Model.

DIAGRAMS OF CAUSAL RELATIONS

Arrows are drawn to each dependent variable from each independent variable that influences it directly, as postulated in the equations of the system.

and as this discrepancy persists, the group will begin to *perceive* that A is unreceptive. This perception in turn, through equation (1.3), reduces the cohesiveness of the group with A. Equation (1.2) was not of importance in the aggregate model because there was no reason to postulate that the members would be highly unreceptive.

Finally, equation (1.4) represents a generalization of (3.5) to take account of the possibility of subgroup formation. In the deviate model, pressure toward uniformity with A depends *both* on the internal cohesiveness of the rest of the group (C) *and* on their cohesiveness with A (C_A), a distinction we did not have to make in the completely aggregative model.

By way of summary we give a diagram in Fig. 4*a* and 4*b* of the causal relations among the variables implied by the two systems of equations. For simplicity we assume change in variables to take place at discrete intervals rather than continuously. Thus, $L_A(t+1)$ is the value of L_A one time interval after t—similarly for C_A, D, and C.

The deviate-member and aggregate models may both be regarded as special cases of a more general implicit model in which aggregates do not appear, but only variables for individual members of the group. One special case (corresponding to our aggregate model) arises when the group behaves in an 'undifferentiated' manner, with similar mechanisms operating for all members. Another arises when the behaviour of one group member (or a small minority) becomes differentiated from the others, while the behaviour of those others can still be aggregated—our deviate-member model. When they are justified empirically it is advantageous to make these simplifying assumptions rather than deal with the more cumbersome general model. Similarly, in treating certain of the empirical studies with our aggregate model, we have ignored particular equations when the mechanisms they represented were not operating in the empirical data being examined. In formulating the deviate-model we added implicitly a mechanism of primary importance that needed to be only implicit in the aggregate model. This does not mean that we are explaining the several studies in terms of *different* theories, but rather that we can conceptualize a general theory wherein the particular phenomena under study

represent special cases. For the present it seems unprofitable to write the more general equations, which would consist of composite sets of equations analogous to those used in the two models, one set per group member. They would be complicated and difficult to handle, and at first glance do not seem capable of yielding fruitful derivations.

IV. COHESIVENESS AND REJECTION

Before concluding, we should like to comment on the cohesiveness variables, C_A and C, that appear in the two models. Cohesiveness is several times defined by Festinger and his colleagues as ' the resultant of all the forces acting on the members to remain in the group '. It plays the role of an intervening variable between (*a*) the various forces that increase or decrease the attractiveness of the group, and (*b*) the behaviours that result from these forces. Its usefulness as an intervening variable depends on whether, in fact, all the various types of positive and negative attraction to the group do produce the same kind of behaviour provided only that their net intensity is the same. The evidence on this is far from conclusive.

When C is defined as the ' resultant of all forces to remain in the group ', the influence of the attractiveness of other groups is ignored. By interpreting C as the ' net attractiveness ', it would seem that a wider array of phenomena could be encompassed in the aggregative model. Only one of Festinger's hypotheses considers the operation of the attractiveness of other groups:

> *Hypothesis* 3*c:* The amount of change in opinion resulting from receiving a communication concerning ' item x ' will decrease with increase in the degree to which the opinions and attitudes involved are anchored in other group memberships or serve important need satisfying functions for the person.

With a ' net attractiveness ' interpretation of $C(t)$, this hypothesis states a relation between ΔD and C via L and U, and is in fact identical with hypothesis 3*b*:

> *Hypothesis* 3*b:* The amount of change in opinion resulting from receiving a communication will increase as the strength of the resultant force to remain in the group increases for the recipient.

That is, if C is the net attractiveness of the group, it can be increased either by increasing the forces to remain in the group or by decreasing the attractiveness of other groups. Hence the variable in 3*b*: ' strength of the resultant force to remain in the group ' is the negative of the analogous variable in 3*c*: ' degree to which the opinions and attitudes involved are anchored in other group memberships '.

In this context, it may be well to call the reader's attention to our interpretation of the cohesiveness variables in the deviate-member model, wherein we regarded negative values of C_A as ' rejection '. This again is an effort to define the variables in the system more parsimoniously. The agreement of the deviate model with Schachter's data indicates that no new ' rejection ' variable, apart from C_A is necessary to account for Schachter's results.

If one wanted to distinguish the attracting forces of other groups from the attracting forces of the group upon which attention is centred, two cohesiveness variables might be used, as C_1 and C_2; the ' net attractiveness ' then being $C_1 - C_2 = C$, or cohesiveness.

There is another respect in which the cohesiveness variable is ambiguous. The consequences of low cohesiveness certainly must be expected to depend upon what alternatives a member sees to remaining in the whole group. There are at least four obvious possible reactions to a discordant group situation: (1) the member may consider the alternative of withdrawing from the group, (2) he may remain but become less sensitive to its attempts to control his opinions and behaviour, (3) he may consider the alternative of ejecting from the group one or more members who cause the discord, and (4) he may consider division of the group into subgroups. Hence, each member's perception of the distribution of group opinions, the possibilities of subgroup formation, his influence upon the group, and his opportunities for membership in other groups all affect the significance of cohesiveness. The predictions of our models all involved implicit assumptions with respect to these perceptions. The variable C would seem to be interpretable as the average intensity of the desire to maintain the group, but with the possible ejection of one or a few members. The variable C_A might be best interpreted as the average desire to retain a particular member of the group. No attempt was made to carry out these refinements. We do wish, however, to call attention to their importance in the further development of models of this kind.

Comment on Mathematization. The utility of mathematics in dealing with phenomena of the kind under discussion here is by no means universally accepted. What advantages do we think are gained by such an undertaking? We have tried, in examining Festinger's

propositions, to illustrate concretely some of the advantages derivable from the construction of a formal mathematical model:

(*a*) Knowing more precisely what mechanisms or structural relations are being postulated, and sometimes calling attention to the need for further clarification of the operational meaning of definitions and statements;

(*b*) Discovering whether certain postulates can be derived from others, and hence can be eliminated as independent assumptions; whether additional postulates need to be added to make the system complete and the deductions rigorous; and whether there are inconsistencies among the postulates;

(*c*) Assisting in the discovery of inconsistencies between the empirical data and the theories used to explain them;

(*d*) Laying the basis for the further elaboration of theory, and to deductions from the postulates that suggest further empirical studies for verification;

(*e*) Aiding in handling complicated, simultaneous interrelations among a relatively large number of variables, with some reduction of the obscuring circumlocutions entailed by non-mathematical language.

In the long run, mathematics will be used in the social sciences to the extent that it provides a sufficiently powerful language of analysis and exposition to justify the time and effort required to use it. Social phenomena have proved sufficiently baffling; whether our inclinations are mathematical or literary, we cannot afford to prejudice what is the 'one best' methodology. Let methods, like substantive theories, prove by actual trial which is the more powerful, and under what conditions.

V. SUMMARY

1. In this paper we have examined certain hypotheses concerned with pressures toward uniformity upon the deviate-members of groups, and have endeavoured to develop them into an integrated, coherent system.

2. We have devized a model containing feed-back mechanisms to provide linkage among such variables as the perceived receptivity of the deviate-member and the desire of the group to retain him in the group.

3. Derivations from the model made it possible to check the system's congruence with empirical findings in an experimental situation. The check confirmed certain assumptions incorporated in the model, but showed also that many of the hypotheses set forth by Festinger and his associates simply were not being tested by the experiments. The mathematical derivations made it possible to predict time paths for certain of the variables and to check these predictions against the data.

4. A comparison was made of the deviate-member model with a homologous model developed for groups without a deviate. The two were found to be special cases of a more general system.

5. Two seemingly independent variables, cohesiveness and rejection, were found to be interpretable as merely opposite ends of a continuum. The new interpretation yielded derivations which were consistent with empirical data, and the consolidation helped increase the parsimony of the model.

APPENDIX. SYSTEMS OF DIFFERENTIAL EQUATIONS

The models with which we have been dealing attempt to explain a set of phenomena in terms of four or five mechanisms, each mechanism being represented by a differential or algebraic equation. Implicit in this method is the assumption that systems can actually be observed, in field or laboratory, whose behaviour is determined by these mechanisms and which are isolated from other systems to a sufficient degree of approximation.

Consider a system of n differential equations in n variables:

$$\frac{dx_i}{dt} = f_i(x_1,..., x_n) \qquad (i = 1,..., n) \qquad (1)$$

Arrange these equations in order of the 'rapidity' of the adjustment processes: (cf. below, p. 101). Let us suppose that we observe the whole system at intervals of τ minutes (or milliseconds, or years) and over a total span of T minutes (or milliseconds, or years). Now if the system starts at time zero

from a disequilibrium position, certain of the variables (the first k, say, which adjust most rapidly) will effectively reach their equilibrium values (within, say, ± 1 per cent.) within τ minutes, and hence before the next observation of the system. For these we shall have at all times (approximately):

$$f_i(x_1,..., x_n) = 0 \qquad (i = 1,..., k) \qquad (2)$$

which gives k algebraic equations to determine the first k variables in terms of the remaining $(n-k)$.

Among the remaining $(n-k)$ variables, some may have such slow adjustment processes that their values at time T are effectively the same (within, say, ± 1 per cent.) as at time 0. Suppose that these are the variables numbered from $(m+1)$ to n. Then the corresponding equations can be eliminated from (1), and the variables treated as constant parameters.

We have left, then, in addition to equations (2), a set of $(m-k)$ differential equations of the form (1), from which we can determine the time paths of the variables $x_{k+1},..., x_m$. We see that by appropriate selection of the time units, τ and T, and of the initial position, $(x_1^0,..., x_n^0)$, of the system we can select out various subsets of the equations (1) for study.[1] Moreover, if we study the behaviour of the system as it moves from various initial positions, we can study the effect of various of the parameters (i.e., the variables $x_{m+1},..., x_n$) upon it.

If the system of (1) is linear, we can make these statements more explicit. The general solution of the system is then:

$$x_i = \sum_{j=1}^{n} a_{ij} e^{\lambda_j t} \qquad (3)$$

where the x's are measured from their equilibrium values, the λ's are complex numbers and the a's are constant. The system is stable if and only if all the λ's have negative real parts. The more negative the real part of λ_k the more rapidly does the term $a_{ik}e^{\lambda_k t}$ approach zero. If we now arrange the terms in order, so that the λ with largest real part comes first, and so on, those adjustment processes will be rapid for which all the a's except the first few in (3) are very small, while those adjustment processes will be slow for which all the a's except the last few in (3) are very small.[2]

[1] This is not, of course, the only way of conceptualizing the compartmentalization of a general system into independent subsystems. For a very important proposal along different lines see W. Ross Ashby, *Design for a Brain* (1952), particularly Chap. XI, XIV-XVIII, and XXIV. The whole problem is related to the question of causal ordering. See H. A. Simon, 'Causal ordering and identifiability'. In T. Koopmans (ed.), *Studies in Econometric Method*, Cowles Commission for Research in Economics, Monograph, XIV, 1953.

[2] In conversation, Dr. John W. Tukey has pointed out that this three-way classification of processes into 'rapid' (high frequency), 'slow' (low frequency), and 'intermediate' (intermediate frequency) is involved in every empirical analysis of time series. The over-all period of observation, T, places a lower limit on the frequencies that can be examined, and the interval of observation, τ, an upper limit. While Dr. Tukey's comments refer to the oscillatory components of time series, and ours to the exponential components the point involved is the same: any given set of observations only permits us to explore a subset of the mechanisms at work, and these are the mechanisms which are neither too slow relative to T, nor too fast relative to τ.

REFERENCES

1. Festinger, Leon (1950). 'Informal social communication.' *Psychological Review*, LVII, 271-282.
2. Schachter, Stanley (1951). 'Deviation, rejection, and communication.' *J. Abnor. Soc. Psychol.*, XLVI, 190-209.
3. Simon, H. A. and Guetzkow, H (1955). 'A model of short- and long-run mechanisms involved in pressures toward uniformity in groups.' *Psychol. Rev.*, LXII, 56-68.

Vol VIII
Part II

The British Journal of Statistical Psychology

November
1955

TEST RELIABILITY ESTIMATED BY ANALYSIS OF VARIANCE[1]

By CYRIL BURT

Historical Note. The idea of determining the 'precision' of an estimated quantity from a series of repeated measurements is due originally to Gauss (*Theoria Combinationis Observationum*, 1809). It was introduced into psychology by Fechner who proposed a *Präzisionsmass im Gaussschen Sinne*, viz., $h = 1/\sigma\sqrt{2}$ (*Elemente der Psychophysik*, 1859, pp. 103 f.). Clarke Wissler was apparently the first to suggest that 'the precision of a test may be estimated by *correlating successive trials*' (*Psychol. Mon.*, III, 1901, p. 60). The correlation so obtained has generally been known as a 'reliability coefficient'[2]: it seeks to answer the question—'if this test were repeated, how far would the results agree?' The traditional way of raising the question would have been to ask: 'if this test were repeated, how much variation would there be in the results?' The suggestion that the 'reliability' of tests and examination marks might be more effectively studied by analysing the variation—with the aid either of factor analysis or of what Fisher has termed the analysis of variance—was, I believe, first put forward in a Memorandum I was asked to prepare for the International Institute Examinations Enquiry[3] in 1934; and a number of theoretical investigations with these newer techniques have since been published.[4] Their adoption for practical purposes has, until the last few years, been comparatively rare.

Definitions. Both in this *Journal* and elsewhere several writers have recently complained that, 'in spite of the vast literature that is available, psychologists still use the term reliability in a loose and ambiguous fashion'.[5] What kind of 'agreement' or 'variation' is to be regarded as a manifestation of 'reliability' or the reverse? We must therefore begin by defining a little more exactly what it is we are seeking to assess. The implicit notion is that of 'error', which is

[1] *Editorial Note.* This paper is intended to form one of a series of elementary 'expository notes' which the Journal Committee has suggested should be published. The formulation of the problem and the deduction of the various algebraic equations are therefore not so rigorous as would be desirable in an exposition designed for more advanced students. Points that might appear technical or even pedantic to the ordinary reader I have, so far as possible, relegated to footnotes. The greater part is taken from the roneo'd *Laboratory Notes on Reliability* (1938) and *Analysis of Variance* (1943) referred to by Mr. Mahmoud in the paper that follows. These notes in turn were amplifications of the Memoranda drawn up for the Examination Enquiry Committee in 1931: (see *Marks of Examiners*, 1936, and *The Marking of English Essays*, 1941).

From the correspondence that followed Mr. Mahmoud's note on the 'assessment of reliability by analysing variance' (this *Journal*, VII, 1954, p. 61) it appeared that many readers would welcome a more accessible account of the methods described in these publications; and the Editor therefore hopes that this paper may serve to explain their theoretical basis and that Mr. Mahmoud's may sufficiently illustrate their practical application.

[2] The substitution of the word 'reliability' arises from the fact that in psychology most of the early experimentalists received their training in Germany. At the turn of the last century, when German writers began to substitute Teutonic terms for French, *Zuverlässigkeit* was proposed as a synonym for *Präzision*. It was, for example, adopted by Krüger and Spearman to designate Wissler's 'test-retest correlation' (*Zeitsch. f. Psychol.*, XLIV, 1906, pp. 48). In his English writings Spearman used the ordinary English translation for *Zuverlässigkeit*, namely, 'reliability'. Cattell and Thorndike use the word 'reliability' in the wider sense, to cover *all* measurements of precision: (cf. E. L. Thorndike, *Mental and Social Measurements*, 1904, chap. X and *passim*).

[3] In preparing the memorandum I was greatly aided by the work and advice of my colleagues in the Statistical Department of University College, particularly Professor R. A. Fisher, Dr. P. O. Johnson, and Dr. J. Neyman; and in revising the notes for teaching purposes I was much indebted to various research students, particularly to Miss M. Howard and Mr. R. W. B. Jackson.

[4] Cf. more particularly P. O. Johnson, 'Tests of Certain Linear Hypotheses and their Application to Some Educational Problems', *Statistical Research Memoirs*, I (1936), 57-93, R. W. B. Jackson, 'The Reliability of Mental Tests', *Brit. J. Psychol.*, XXIX, 267-87, and G. A. Ferguson, *The Reliability of Mental Tests* (1941) and refs.

[5] Cf. A. S. C. Ehrenberg, 'The Reliability of Repeated Measurements', *Bull. Brit. Psychol. Soc.* No. 23 (1954), 3 f., and this *Journal*, VI, 41 f., and A. Heim, *The Appraisal of Intelligence* (1954), 67 f.

itself a somewhat elusive concept. For statistical purposes a working definition may be derived as follows.[1]

In psychological research, the investigator continually requires to measure as accurately as possible certain aspects or types of behaviour (which we will call ' traits ') manifested by certain individuals. The most elementary observation that he can make, therefore, will consist in using (i) some experimental device, which we call a *test*, to elicit and assess the specified trait (ii) in some particular *individual* (iii) on some particular *occasion*. In any such experiment, the particular test, the particular person, and the particular occasion will each be specimens selected from a whole universe or population of such tests, persons, and occasions. We are thus concerned with variations in three distinct dimensions (see Fig. 1).

Accordingly, let x_{ijk} denote the observed measurement actually obtained for the ith individual with the jth test on the kth occasion ($i = 1,\ldots, N$; $j = 1,\ldots, n$; $k = 1,\ldots, m$); let ξ_{ij} be the ' true ' measurement, i.e., the measurement we should obtain if we possessed a perfectly efficient ' test '[2]; and let ϵ_{ijk} denote the error of the measurement actually obtained. Then, adopting algebraic addition as the simplest way of relating these quantities, we may write

$$x_{ijk} = \xi_{ij} + \epsilon_{ijk} \quad \text{or} \quad \epsilon_{ijk} = x_{ijk} - \xi_{ij}, \qquad \text{(i)}$$

and we may take the latter equation as a formal definition for the ' error ' of a single observation.

This formulation involves two unknowns, but only one equation. To obtain equations that are soluble, the most obvious procedure would be (i) to secure measurements for the same person with the same test on a number of *different occasions*. In practice this may not always be feasible. Often it would be far more convenient (ii) to obtain measurements from the same person on the same occasion with a number of *different* ' *tests* ' (e.g., different problems or ' items ' in a composite group test), or (iii) to obtain measurements on the same occasion with the same test from a number of *different persons*, or finally (iv) to use some combination of these various alternatives. We could then extract an estimate, not for the amount of *each* error (ϵ_i), but for the general tendency to error : e.g., with the first procedure

$$\bar{e}_{ij} = \frac{1}{m}\sum_{k=1}^{k=m} |x_{ijk} - \bar{x}_{ij}| \quad \text{or} \quad s_e = \sqrt{\left\{\frac{1}{m-1}\sum_{k=1}^{k=m}(x_{ijk} - \bar{x}_{ij})^2\right\}},$$

where m as before denotes the number of occasions on which the test is applied.

Types of Error. Following Gauss and Fechner, psychologists commonly distinguish two main kinds of error : (i) systematic or ' constant ' errors, which are more or less predictable and so avoidable ; (ii) chance or ' random ' errors, which are unpredictable and unavoidable. Here we shall be

[1] As Dr. Ehrenberg points out, most definitions involve a number of dubious assumptions (*loc. cit.*, p. 41) : some of them indeed I myself believe to be unnecessary. However, Dr. Ehrenberg's criticisms of Professor Gulliksen's assumptions in *The Theory of Mental Tests* (1950) appear unduly severe : but cf. Dr. Guttman's review and Dr. Gulliksen's reply (*Psychometrika*, XVIII, 123-34).

[2] On the whole this simple definition of a 'true measurement ' seems to me the best : I have phrased it so as to cover ' true measurements ' that are purely hypothetical quantities as well as those (if any) which are actual quantities, e.g., Tom's innate general ability (which by hypothesis is unchanging) as well as, say, Tom's visual acuity or his efficiency in arithmetic. But the popular notion that this or that characteristic has in fact a ' true ' measurement, if we could only find a perfect instrument for discovering it, is to my mind a highly dubious ' metaphysical ' concept (' metaphysical ' in the pejorative sense in which the word is used by the logical positivists). In psychology more particularly almost every specifiable trait is an unstable, fluctuating characteristic. Its measurement involves the determination of a region rather than a point. We can, if we like, define the particular point which, as it were, forms the centre of gravity about which the momentary values appear to vary. But *these internal variations do not form part of the ' unreliability ' of our test or other metrical technique.*

To circumvent what I have called the metaphysical difficulty, it is tempting to fall back on some kind of operational definition : the ' true ' measurement we may say is merely ' the limiting value towards which the mean of a number of measurements progressively tends as the number is indefinitely increased ', or (in more modern terminology) ' the mathematical expectation over the whole universe of trials ' : this would be equivalent to defining it by some such equation as $\xi = \mathscr{E}(x) = \frac{1}{m}\sum_{k=1}^{k=m} x_{ijk}\ (m \to \infty)$. But there are pitfalls in this definition. In particular, we shall be liable to include in the ' error ' the variations due to changes in the trait itself as well as those due to the imperfections of the method of measurement. Hence I should prefer to treat this equation as an *inference* from the definition I have given above. For the general approach adopted in this section see F. N. David, *Probability Theory for Statistical Methods* (1949) chap. X ; in accordance with the continental practice followed by Dr. David and others, I use $\mathscr{E}$ rather than E for such operations as $\frac{1}{m}\sum_{k}^{m}$.

concerned almost exclusively with random errors. A random effect is usually described as one that results from a very large number of very small independent causes, about which nothing is known that could serve as a basis for prediction. I should prefer to follow Keynes, and define both 'randomness' and the related notion of 'independence' in terms of 'irrelevance'. From these definitions[1] several important corollaries can then readily be deduced. If the error e is a random variable in the sense defined, then as m increases,

$$\text{(i)}\ \underset{k}{\mathscr{E}} e_k \equiv \frac{1}{m}\sum_{k}^{m} e_k \ \rightarrow\ 0, \text{ and hence}^2\ \text{(ii)}\ \underset{k}{\mathscr{E}} x \equiv \bar{x} \ \rightarrow\ \xi\ ;$$

$$\text{(iii)}\ r_{e_j e_{j'}} \equiv \underset{k}{\mathscr{E}}(e_{jk}\, e_{j'k}) \ \rightarrow\ 0\ ;\ \text{similarly,}\quad \text{(iv)}\ r_{\xi e} \rightarrow 0\ ;$$

and (v) with measurements of the kind with which the psychologist has to deal, the distribution of the errors will tend to the normal. With suitably defined conditions these corollaries will also hold, not only when the expectations refer to data obtained on a number of different occasions, but also when they refer to a number of different persons; and most of them can be subjected to an empirical check.

Definition of Reliability. (*a*) *In Terms of Variation.* Since we frequently need to compare the reliability of different tests, we cannot leave the measurements of error in units specific to each test or trait. The time-honoured device is to take as unit the standard deviation of the general population, or (by preference, since it is additive) the square of the standard deviation, i.e., the variance. This can be estimated by choosing a random sample of the population, and calculating $s_x^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i-\bar{x})^2$. On comparing the two variances it would then seem possible, on intuitive grounds, to infer that, when the variance of the measurements for a single individual (say the ith, $s_{e_i}^2$) becomes as large as the variance for the entire sample of different individuals (s_x^2), the test used will be of no practical value whatsoever: for the whole object of such a test is to *distinguish* the ability as measured for any given individual from the abilities of the rest. On the other hand, if the variance of the measurements for a single individual were zero, we should pronounce our method of measurement to be entirely free from error. The unreliability of the test will then be measured by the ratio of the individual variance to the total variance treated as unity. And for any given test j we may calculate the quantity

$$r_v = 1 - s_{e_i}^2/s_x^2,$$

which will evidently be 0 when $s_{e_i}^2 = s_x^2$, and 1 when $s_{e_i}^2 = 0$; and we may regard this as estimating on a scale of 0 to 1 the precision or reliability of the test. It will be seen that this approach treats reliability as *an invariant property of the test* (where 'test' means not merely the test-material but the entire process of measuring the specified trait for the specified population).

So far as the assumptions we have made hold good with the sample of persons actually measured, it will be a matter of indifference whether our sample of trials is obtained by repeated testing of one and the same typical individual or by testing a number of different individuals[3]; for, by corollary

[1] *Treatise of Probability*, pp. 55, 120, 291, 412; cf. also M. G. Kendall, *Advanced Theory of Statistics*, I, p. 171 f. and refs. p. 184, and David, *loc. cit. sup.* In the notation adopted by Keynes, X_1 is said to be irrelevant to X_2/h, if $X_2/X_1h = X_2/h$, i.e., if the probability of X_2 on the evidence X_1h is the same as its probability on the evidence h alone. Now let $C(X)$ mean 'the variable X is a member of the class C', and let $\phi(X)$ mean 'X has the characteristic or the value ϕ'; then a is said to be a 'random' member of the class C, if the statement 'X is a' is irrelevant to the probability $\phi(X)/C(X).h$, or (better perhaps) if $\phi(a)/C(a).h = \phi(b)/C(b).h$, where $C(b).h$ contains no information about b, except that it is also a member of the class C. The 'theorem of independence' states that if $X_2/X_1h = X_2/h$, then $X_1/X_2h = X_1/h$, and X_1/h and X_2/h are said to be independent. 'Statistical independence' is commonly defined in terms of statistical frequency functions: if the variation of X_1 is irrelevant to that of X_2, i.e., if the distribution of X_2 is the same whether X_1 varies or not, then X_1 and X_2 are said to be independent; and in particular $f(X_1, X_2) = f_1(X_1)f_2(X_1)$ where $f(X)$ as usual denotes the 'frequency function' of X: (cf. Kendall, *loc. cit.*, p. 21, eqn. 1.22). Furthermore, $\mathscr{E}X_1X_2 = (\mathscr{E}X_1)(\mathscr{E}X_2)$ or $\mathscr{E}(X_1-\mathscr{E}X_1)(X_2-\mathscr{E}X_2) = 0$: in other words, the covariance is zero. This provides a useful test of independence: unless the covariance of X_1 with X_2 is statistically non-significant (and not always even then) we cannot justifiably assume that X_1 and X_3 are statistically independent.

[2] To avoid complicating the argument in the text I have omitted certain qualifications which a more rigorous statement would require; e.g., even supposing that the expectation of the sample mean converges equal to the population mean (p. 104, n. 2), this would not of itself *imply* that the sample mean converges in probability to the population mean: still, under conditions which can nearly always be assumed in practice for data of the type here considered, it will converge.

[3] At this point, I think, my treatment diverges from that adopted by Dr. Guttman ('Basis for Analysing Test-Retest Reliability', *Psychometrika*, X, 1945, pp. 255-82—an article which every student should study). Dr. Guttman restricts error variance to the variance of measurements for the ith individual, taken about the

(iv), there should be no relation between the size of the errors and the size of the true measurements. But in practice, if there *were* a possibility of some such relation, it would plainly be safer to draw our sample of errors from the whole range over which the test is to be applied. On these grounds, as well as those of ordinary convenience, it will be better to measure *all* the individuals at least twice, and then substitute the *mean* of the error variances in the numerator and the variance of the observed *means* for the different individuals in the denominator thus obtaining an alternative formula :

$$r_v = 1-\bar{s}_e^2/s_x^2 \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(ii)}$$

Now, even if the test had no diagnostic value, the means of the measurements for the different individuals would not be exactly the same. We should in fact expect their variance to be approximately equal to the error variance. Hence, when the variance of the observed means exceeds the error variance, we may reasonably attribute this excess to differences in the true measurements : in fact, it follows algebraically from eqn. (i) that, if (as we have assumed) the true measurements and the errors are uncorrelated, $s_x^2 = s_\xi^2+\bar{s}_e^2$. Accordingly, we may estimate the variance of the 'true' measurements from the difference between the two. Thus, in practice, the formulae just reached are equivalent to saying that the reliability of the test can be defined as that proportion of the *total* variation observed which is attributable to the variation of the *true* measurements : i.e.,

$$r_v = \frac{s_x^2-s_e^2}{s_x^2} = \frac{\hat{s}_\xi^2}{s_x^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(iii)}$$

where the circumflex accent denotes that the variance of ξ is merely an estimate.

(b) *In Terms of Agreement.* If the *differences* between the measurements obtained for one and the same quantity provide an indication of their *unreliability*, it is natural to treat their *agreement* as an indication of *reliability*. At its simplest this agreement can be assessed by computing the correlation between the two series of measurements obtained with identical forms of a given test (j and j'), applied under identical conditions, namely,

$$r_{xx'} = \frac{\sum_i x_{ij}x_{ij'}}{\sqrt{(\sum_i x_{ij}^2 \sum_i x_{ij'}^2)}} \qquad (i = 1, 2,\ldots, N).$$

Now $\sum x_{ij}\, x_{ij'} = \sum (\xi_i+\epsilon_{ij})(\xi_i+\epsilon_{ij'}) = \sum \xi_i^2+\sum \xi_i\epsilon_{ij}+\sum \xi_i\epsilon_{ij'}+\sum \epsilon_{ij}\epsilon_{ij'}$, and this (by corollaries ii and iii above) tends to $\sum \xi_i^2$ as N increases. Similarly $\sum x_{ij}^2 \to \sum \xi^2 + \sum \epsilon_{ij}^2$ and $\sum x_{ij'}^2 \to \sum \xi_i^2+\sum \epsilon_{ij}^2$. Moreover, since by hypothesis the two forms of the test are identical, $\sum \epsilon_{ij} = \sum \epsilon_{ij'}^2 = \sum \epsilon_i^2$. Hence, under the conditions assumed,[1] we may write

$$r_{xx'} = \frac{\sum \xi_i^2}{\sum \xi_i^2+\sum \epsilon_i^2} = \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\varepsilon^2} \quad . \quad . \quad . \quad . \quad . \quad \text{(iv)}$$

or, substituting estimates,

$$= \frac{s_x^2-s_e^2}{s_x^2} = 1-\frac{s_e^2}{s_x^2} = r_v. \quad . \quad . \quad . \quad . \quad . \quad \text{(v)}$$

In passing we may note that, since

$$r_{xx'} \to \frac{\sigma_\xi^2}{\sigma_x^2} \quad \text{and} \quad \frac{\sigma_\xi}{\sigma_x} = \frac{\sigma_\xi^2}{\sigma_\xi\sigma_x} = \frac{\sum \xi x}{\sqrt{(\sum \xi^2 \sum x^2)}} = r_{\xi x},$$

we may take $\sqrt{r_{xx'}}$ as an estimate of $r_{\xi x}$, the 'index of reliability' (Kelley, *Statistical Method*, 1923, p. 206, eqn. 160). Hence, unless the 'criterion' (i.e., the independent estimates obtained for the

---

mathematical expectation and over 'an indefinitely large *universe of trials*'. This gets rid of many of the assumptions otherwise required. Now, although an economy of assumptions is desirable in *theoretical* work, in *practical* work it seems in general wiser to retain them, and so far as possible, to verify to what extent they are fulfilled.

[1] For eqn. iv to be valid a less stringent requirement would serve. As the proof implies, the necessary and sufficient condition is merely that the two columns of measurements to be correlated are homoscedastic.

trait) is itself as unreliable as the test, it is quite erroneous to state, as is so often done, that 'the validity of a test can never exceed its reliability'.[1]

Equation iv expresses in its simplest form a mode of approach first suggested by Galton, namely, an analysis of variation into independent components or 'factors': if the total variance, σ_x^2, can be divided into two independent and additive portions, one due to the factor *common* to both variables (σ_c^2) and the other due to the factors *specific* to either variable (σ_e^2), then the amount of agreement between the two sets of measurements can be expressed by the ratio of the common factor variance to the total variance.[2]

Owing to the fact that the non-statistical psychologist always finds it easier to think in terms of correlations, he has, as a rule, nearly always preferred to compute a kind of 'self-correlation' by applying the familiar product-moment formula. But the expression we have reached (eqn. v) shows that the same value can be got by adopting an analysis of variance instead of the more usual procedure. Such an analysis will have several obvious advantages: it will enable us to check the underlying assumptions, to obtain the same coefficient by a more direct calculation, to construct modified forms of the coefficient where required, to adapt the concept to cases where the test has been applied only once or generalize it to cases where the test has been applied more than twice, and to extend the treatment to investigations in which all three sources of variation—persons, tests, and trials—and even others as well, can be compared simultaneously.

I. THE RELIABILITY OF A SIMPLE TEST

Illustrative Example. The procedure proposed can be most simply explained if we start with an actual example. Let us suppose that, by means of a booklet comprising five sets of questions, we have measured the intelligence of six pupils, chosen at random, and have applied the test on two successive occasions. The detailed marks[3] obtained at each of the trials are set out in Table I. The two rows of figures near the bottom of the table, which give the averages for the five subtests, will form the 'final marks' for the first and second trial respectively. The reader may, if he likes, think of the figures as 'intelligence quotients'. To deal with the problem in its simplest form first, we shall for the moment ignore the data for the constituent subtests, and merely ask what is the reliability of the composite test, judged by the 'final marks'.

If identical tests, or two equivalent (that is, interchangeable) forms of the same test, have been used for the two trials, we may expect that both the means and the standard deviations will be approximately equal on each occasion. Since the occasions are themselves samples, we cannot expect *exact* equality; but what small or non-significant discrepancies there may be can be treated as an incidental result of the unreliability. Accordingly, the best estimate of the mean for the entire sample will be the mean of the two averages obtained for the whole group on the two occasions (101.1). Similarly the best estimate for the standard deviation will be derived from the average of the two square-sums taken about this mean. The coefficient of correlation obtained from estimates so computed is termed the intra-class correlation; and, as I have argued elsewhere, it appears to be the most suitable form to adopt for this purpose. It has the merit of being based on $2N-1$ instead of $N-1$ degrees of freedom, and thus provides a more accurate estimate than the ordinary product moment coefficient. On this basis, therefore, the formula will be (R. A. Fisher, *Statistical Methods for Research Workers*, 1935, p. 199)

$$r_{int} = \frac{\sum\{(x-\bar{x})(x'-\bar{x})\}}{\frac{1}{2}\{\sum(x-\bar{x})^2+\sum(x'-\bar{x})^2\}} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(vi)}$$

[1] Several writers have argued that, if their tests have been shown to be valid, 'they must therefore be reliable' (cf. H. J. Eysenck, this *Journal*, VI, 1953, pp. 44). Some of Dr. Eysenck's tests, for example, have theoretical validities of over 0·70—admittedly a high figure for tests of personality; but a test with a theoretical validity of 0·70 might have a reliability of only $0{\cdot}70^2 = 0{\cdot}49$, which is not a satisfactory value for a reliability coefficient.

[2] It is instructive to compare the ratio here reached with the expression for 'information communicated' in arguments from information theory. To take the simplest case, let the 'true message' to be communicated be denoted by ξ and the obscuring 'noise' by ϵ, and let the message as received be denoted by $x = \xi+\epsilon$: then, as Wiener proves, the information gained can be measured by $\frac{1}{2}\log_2 \frac{\sigma_\xi^2+\sigma_\epsilon^2}{\sigma_\epsilon^2}$. The formula can be generalized to meet the more complex cases where a multiplicity of variables and common factors are involved (cf. N. Wiener, *Cybernetics*, 1948, and Burt, this *Journal*, IV, 1951, pp. 193 f.). In my fuller *Notes* I endeavoured to show that a more satisfactory derivation of the formulae for reliability could be obtained if the basic definitions and postulates were expressed in terms of information theory. However, this line of approach seemed too technical for adoption here.

[3] The detailed figures are taken from K. Mather, *Statistical Analysis* (1943), p. 73, Table 12. The reason for the choice will be explained in a moment (p. 111). To save space the tables give values to one decimal place only: the calculations are based on two or more.

where x denotes the first set of N measurements and x' the second set, and $\bar{x}$ denotes the mean of the entire series. The formula, it will be seen, departs from the more familiar version by substituting the arithmetic mean of the two variances for the geometric.

Applying it to the average marks obtained by the several persons at the two successive trials (the ' final marks ' shown in the last two lines but one of Table I), we obtain

$$r_{int} = \frac{1052{\cdot}85}{\frac{1}{2}\{3945{\cdot}81+2436{\cdot}84\}} = 0{\cdot}3299.$$

With the same data the ordinary inter-class correlation would be $r = 0{\cdot}5288$. The discrepancy, here unusually large, is due mainly to the rather big difference between the means for the two trials.

The Coefficient as a Ratio of Variances. The same value can be reached by calculating the three appropriate variances from the square-sums (i) of the observed values, (ii) of the means for the several persons, and (iii) of the deviations about these means.

TABLE I. UNSTANDARDIZED MEASUREMENTS

Test	Occasion	John	James	Hugh	George	Ralph	Henry	Average
M	1	81·0	146·6	82·3	119·8	98·9	86·9	102·58
M	2	80·7	100·4	103·1	98·9	66·4	67·7	86·20
M	Av.	80·8	123·5	92·7	109·3	82·6	77·3	94·39
S	1	105·4	142·0	77·3	121·4	89·0	77·1	102·03
S	2	82·3	115·5	105·1	61·9	49·9	66·7	80·23
S	Av.	93·8	128·7	91·2	91·6	69·4	71·9	91·13
V	1	119·7	150·7	78·4	124·0	69·1	78·9	103·47
V	2	80·4	112·2	116·5	96·2	96·7	67·4	94·90
V	Av.	100·0	131·4	97·4	110·1	82·9	73·1	99·19
T	1	109·7	191·5	131·3	140·8	89·3	101·8	127·40
T	2	87·2	147·7	139·9	125·5	61·9	91·8	109·00
T	Av.	98·4	169·6	135·6	133·1	75·6	96·8	118·20
P	1	98·3	145·7	89·6	124·8	104·1	96·0	109·75
P	2	84·2	108·1	129·6	75·7	80·3	94·1	95·33
P	Av.	91·2	126·9	109·6	100·2	92·2	95·0	102·54
Final Marks	Av. 1	102·8	155·3	91·8	126·2	90·1	88·1	109·05
Final Marks	Av. 2	83·0	116·8	118·8	91·6	71·0	77·5	93·13
Final Av.		92·4	136·0	105·3	108·9	80·6	82·8	101·09

Expressing the average marks (near bottom of Table I) as deviations about the final mean, we obtain the following figures :

								Square-Sums	Divisor	Variance
(i)	Av. 1	+ 1·7,	+54·2,	− 9·3,	+25·1,	−11·0,	−13·0	6382·7	12	531·9
	Av. 2	−18·1,	+15·7,	+17·7,	− 9·5,	−30·1,	−23·6			
(ii)	Means	− 8·2,	+34·9,	+ 4·2,	+ 7·8,	−20·5,	−18·3	2122·1	6	353·7

And the deviations from these means will be

(iii)	for Av. 1	+ 9·9,	+19·3,	−13·5,	+17·3,	+ 9·5,	+ 5·3	2138·5	12	178·2
	for Av. 2	− 9·9,	−19·3,	+13·5,	−17·3,	− 9·5,	− 5·3			

Dividing the square-sums by the number of items summed, we obtain for the total variance 531·9, for the variance of the means 353·7, for the variance of the individual deviations 178·2. Had the test no diagnostic value, we should expect the observed means to show a random variance, similar

to the variance of the individual deviations. We may therefore take 353·7−178·2 = 175·5 as our estimate of the variance of the ' true ' measurements ; and thus obtain for our estimate of the reliability (in accordance with equation (iv))

$$r = \frac{\sigma_\xi^2}{\sigma_\xi^2+\sigma_\varepsilon^2} = \frac{175{\cdot}5}{531{\cdot}9} = 0{\cdot}3299,$$

the same figure as we reached by the method of correlation.

Analysis of Pooled Variance. For practical computation the foregoing procedure is needlessly cumbersome. By far the quickest method will be to compute, direct from the observed values in the usual way, the square-sums needed for a simple analysis of variance.[1] This will at the same time enable us to test the statistical significance of the coefficient so computed. The complete analysis is shown in Table II. Note that, since the means are, as it were, latent in *both* sets of measurements (that for the first trial and that for the second), the sum of square-sums ' between means ' must be entered as 2×2122·1.

TABLE II. ANALYSIS OF VARIANCE

for Two Applications of a Composite Test

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio	Probability
Between means for persons	$Q_p = 4244{\cdot}2$	$(N-1) = 5$	$V_p = 848{\cdot}8$	2·38	<0·05
Within measurements for persons	$Q_r = 2138{\cdot}5$	$N(m-1) = 6$	$V_p = 356{\cdot}4$		
Sum	$Q_s = 6382{\cdot}7$	$(Nm-1) = 11$	$V_s = 1205{\cdot}2$		

The value for the intraclass correlation can now be obtained from the ratio of the *difference* of the two component square-sums to their *sum*.[2] We have in fact

$$\text{(biased) } r_{int} = \frac{Q_p-Q_r}{Q_p+Q_r} = \frac{4244{\cdot}2-2138{\cdot}5}{4244{\cdot}2+2138{\cdot}5} = \frac{2105{\cdot}7}{6382{\cdot}7} = 0{\cdot}3299.$$

Unbiased Estimate. In point of fact, however, the formulae so far given for the required correlation yield an estimate which is biased,[3] and in general too low. The reader, indeed, may have noticed that a moment ago, to obtain the three variances, I divided by ' the number of items summed '. With Fisher's method of carrying out analysis of variance, we obtain the ' mean squares ' by dividing by the ' degrees of freedom '. And, in accordance with this principle, we can correct for the bias by using the difference and sum of the estimated mean-square in place of the difference and sum of the square-sums. We then have

$$\text{(unbiased) } r_{int} = \frac{V_p-V_r}{V_p+V_r} = \frac{848{\cdot}8-356{\cdot}4}{848{\cdot}8+356{\cdot}4} = \frac{492{\cdot}4}{1205{\cdot}2} = 0{\cdot}4086.$$

The correction is here rather large ; with bigger samples it is usually much smaller.

To test the significance of the value thus found, we calculate a variance-ratio in the usual way (2·38) ; and it will be seen that, if we had no other details about the results of our test, we should be led to infer that the differences between the means for the various persons were statistically non-significant. This inference, however, depends on the assumption that the ' mean square within persons ', as given in the table, provides the best available estimate for the error of measurement. But in the original table of detailed marks we have far fuller data for making such an estimate which we have not yet used ; and this further information, as we shall shortly discover, suggests that much of what is here included in the estimate of ' error ' is really of the nature of ' constant error ' rather than of ' random error ', and consists largely of predictable variations due to peculiarities of the subtests and of the different occasions. Let us therefore now consider how far these irrelevant variations can be isolated or allowed for by undertaking a complete analysis of all the detailed marks.

[1] If, however, the observed values have already been reduced to deviations about the final mean, it will be quickest just to square (*a*) their sums and (*b*) their differences, and then sum the squares : this gives us 8288·4 and 4277·0—exactly twice the square-sum calculated in the usual way, i.e., $2Q_p$ and $2Q_r$, which we then insert in a formula analogous to that given in the next paragraph.

[2] See *Factors of the Mind*, p. 275, eqn. i, putting $k = m = 2$.

[3] See Fisher, *loc. cit.*, p. 213, and *Factors of the Mind*, p. 275, eqn. ii.

II. THE RELIABILITY OF A COMPOSITE TEST

A. UNSTANDARDIZED MARKS

The Three-Way Analysis of Variance. When a composite test or battery has been applied to a group of testees on two or more occasions, then, as we have seen, we are really faced with three[1] main sources of variation : (1) the differences between the *persons* tested, (2) the differences between the *tests* employed, and (3) the differences in the conditions obtaining on various *occasions* when the tests are applied. Moreover, each of these main causes may *interact* with one another, so that, in all,

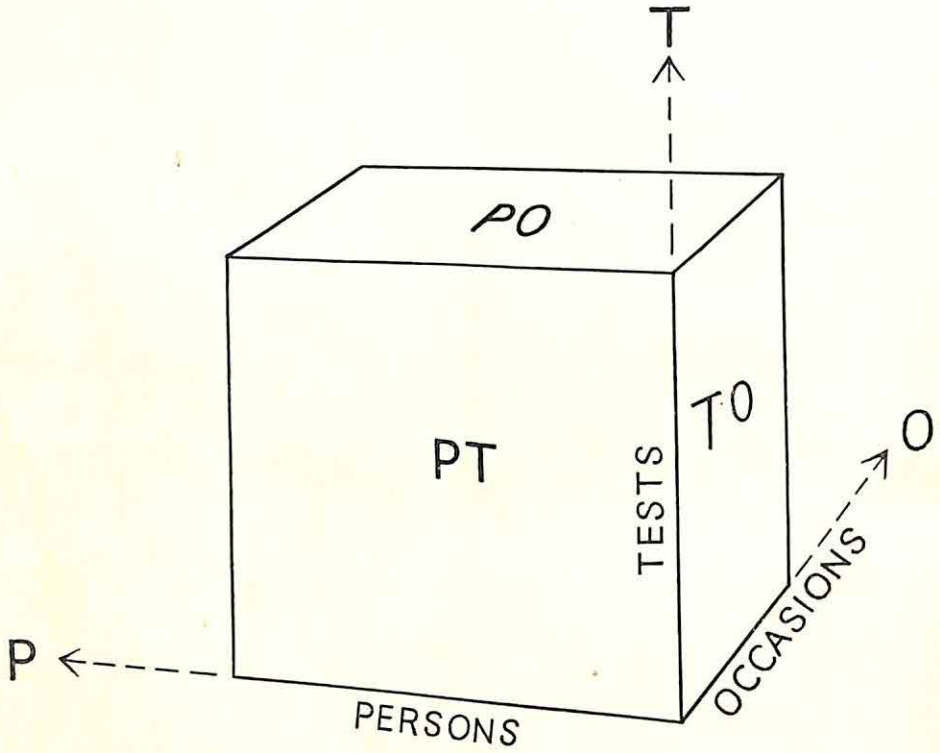


FIG. 1—Analysis of Variance with Three-Way Classification.
Main Dimensions of Variation and Interactions.

the complete analysis will include seven possible sources of variation, namely, three main sources, three first-order interactions, and one second-order interaction which will be used as an estimate of error.[2]

[1] In investigations into ' marks of examiners ' there are frequently *four* main sources of variation : viz. the differences between (i) the question-papers or subjects set (e.g., different topics chosen for an English essay), (ii) the persons examined, (iii) the persons marking the scripts, (iv) the different occasions on which the examination is taken or the scripts marked. As investigations for the International Institute Examinations Enquiry Council showed, an adequate research would involve a specially planned experiment, adopting the method of the ' Latin Square '. For an illustration of this further method, cf. C. Burt and R. B. Lewis, *Brit. J. Educ. Psychol.*, XVI, 1946, pp. 116-32 : the investigation is reproduced by H. M. Walker and J. Lev in the introduction to their chapter on ' Analysis of Variance with Two or More Variables of Classification ' (*Statistical Inference*, 1953, pp. 348-86) to which the research student may refer for further guidance.

[2] For purposes of class-instruction I find it useful to illustrate these various possibilities by geometrical diagrams or models similar to those of Stern. In this case it will be a rectangular parallelepiped : the 3-dimensional solid represents the sample ; the three axes of co-ordinates the three main sources of variation (the indefinite ' universes ' of persons, tests, and occasions) ; the three coordinate planes the 1st-order interactions ; and the volume the 2nd-order interactions or ' error '. Cf. Figure 1.

Our first task therefore is to analyse the total variance (or rather the total square-sum) into the contributions resulting from these various sources. The procedure in this part of our work will be identical with that described in most textbooks of statistics.[1] To save repeating the algebraic proofs involved and the detailed working instructions, the example I have chosen (Table I) consists of a set of data already fully analysed in what is perhaps the most illuminating discussion of the subject for the student of psychology. The student who wishes to follow the actual computations step by step will find them fully set out in the reference given.[2] The final results are shown in Table III.

TABLE III. ANALYSIS OF VARIANCE: UNSTANDARDIZED MEASUREMENTS

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio	Probability
Main Effects					
P. Persons	(Q_p) 21220·9	5	4244·2	30·5	<0·001
T. Tests	(Q_t) 5310·0	4	1327·5	9·5	<0·001
O. Occasions	(Q_o) 3798·5	1	3798·5	27·3	<0·001
First Order Interactions					
PT. Persons and Tests	(Q_{pt}) 4433·0	20	221·7	1·6	0·200-0·100
PO. Persons and Occasions	(Q_{po}) 6893·9	5	1378·8	9·9	<0·001
TO. Tests and Occasions	(Q_{to}) 291·8	4	73·0	—	—
Second Order Interaction				—	—
PTO. Residual	(Q_{pto}) 2784·2	20	139·2		
Sum (Q_s)	44732·3	59			

Main Sources of Variation. There are three preliminary problems to be considered: what evidence do the data afford for differences (i) in the general ability of the persons as tested by the battery as a whole, (ii) in the severity or difficulty of the several subtests, and (iii) in the conditions under which the tests are applied on this occasion or that? The answers are clear. The effects of all three main sources of variation are fully significant.

Of the three variance-ratios, that due to differences between the 'persons' tested is the largest, and now proves to be fully significant. These of course are the differences which our tests were designed to measure; and it is therefore encouraging to find that, even with so small a sample, the investigation yields results that are statistically significant.

The variance ratio due to differences between the two occasions is almost as large. With actual data we might infer that the difference was due to the fact that the second battery is slightly harder, or possibly to some unfavourable circumstance affecting the conditions of the second trial or again to some change in the pupils themselves: had the second mean been higher, we should doubtless have argued that familiarity or practice had improved their general performance; as it is, we might wonder whether the loss of novelty has produced a decline in interest and keenness.

The variance-ratio due to differences between the subtests is far smaller. Here the differences might be due either to differences in their difficulty or to differences in the method of marking. In an actual research such results would indicate that the tests were still capable of considerable improvement.

The Interactions. The interaction between persons and tests can hardly be regarded as significant; but the interaction between persons and occasions is fully significant; while in the case of the interaction between tests and occasions the variance turns out to be smaller than the residual variance and therefore entirely devoid of significance.

Students of psychology usually find the term 'interaction' obscure. Let us therefore glance in passing at the detailed figures which it here denotes. Consider first the interaction between persons and tests. To compute it we begin by averaging the pair of marks (one for each trial) obtained by

[1] E.g., M. G. Kendall, *The Advanced Theory of Statistics* (1946), II, pp. 187 f., K. Mather, *op. cit.*, pp. 72 f., or H. M. Walker and J. Lev, *op. cit.*, pp. 363 f.

[2] K. Mather, *Statistical Analysis in Biology*, with a Foreword by R. A. Fisher (1943), pp. 72-79 (Example 7). Mather's data were in fact agricultural: they might be described as the results of testing the crop-fertility of 6 different soils by using 5 different kinds of seed, sown on two successive occasions. But to make the procedure more readily intelligible to the student of psychology, I have imagined (as stated in the preceding section) that they represent the results of 5 mental tests. A simple illustration of a three-way analysis of variance with *actual* tests was given in an earlier paper in this *Journal* (I, p. 10, Table IV).

each boy in each of the tests. As there are 6 boys and 5 tests, we get 30 averages. We then reduce these averages to deviation form by columns and by rows : in other words, we subtract first the means for the several tests and then the means for the several persons (or vice versa). The final result is the ' doubly-centred measurement matrix ' set out in Table IV. It will be seen that every row and every column adds up to zero. On squaring the figures and taking the sum, we find that the square-sum for these 30 values is 2216·5. But the original mark sheet contained 60 values ; and there each boy's average mark for each test is incorporated in the initial data twice over, once for the first trial and once for the second. In fact Table IV only represents half the interaction, namely, that for the first trial : the other half, that for the second, would be exactly the same. Over the whole table, therefore, the interaction between tests and boys contributes $2 \times 2216 \cdot 5 = 4433 \cdot 0$ to the total square-sum ; and this, it will be seen, is the value inserted in Table III above.[1]

TABLE IV. INTERACTION BETWEEN PERSONS AND TESTS

Person	John	James	Hugh	George	Ralph	Henry	Total
Test M	− 5·34	− 5·84	− 5·91	7·15	8·79	1·15	0·00
S	10·92	2·67	− 4·16	− 7·30	− 1·15	− 0·98	0·00
V	9·06	− 2·69	− 5·95	3·11	4·25	− 7·78	0·00
T	−11·55	16·45	13·18	7·14	−22·07	− 3·15	0·00
P	− 3·09	−10·59	2·84	−10·10	10·18	10·76	0·00
Total	0·00	0·00	0·00	0·00	0·00	0·00	0·00

Sum of squares = 2216·5; $2 \times 2216 \cdot 5 = 4433 \cdot 0$.

What does such a table mean? Inspection of the figures suggests that both tests and boys tend to fall into distinguishable types. James, Hugh, and George appear to do better than the rest at the T-test, and worse at the P- and M-tests ; John, Ralph, and Henry, on the other hand, tend to do worse at the T- and S-tests, and better at the P- and M-tests : there are, however, marked exceptions. This doubly-centred matrix of measurements is in fact equivalent to the table of residuals from which, in an ordinary factor analysis, the bipolar factors would be derived after the general factor had been removed (though in practice such factors would be calculated from the residual correlations, not from the residual test measurements). Indeed the only way of deciding how precisely the boys or the tests should be classified would be to carry out the appropriate factorization. A doubly centred matrix like that we have just computed must have a rank of either $(n-1)$ or $(N-1)$, whichever is smaller. Here, therefore, had the sample of persons been large enough to justify such a procedure, we could have extracted as many as *four* bipolar factors.[2]

TABLE V. INTERACTION BETWEEN PERSONS AND OCCASIONS

Person	John	James	Hugh	George	Ralph	Henry	Total
1st Trial	1·97	11·33	−21·50	9·30	1·56	− 2·66	0·00
2nd Trial	− 1·97	−11·33	21·50	− 9·30	− 1·56	2·66	0·00
Total	0·00	0·00	0·00	0·00	0·00	0·00	0·00

Sum of squares = 1378·8; $5 \times 1378 \cdot 8 = 6893 \cdot 9$.

The interaction between persons and occasions is fully significant. The detailed figures are shown in Table V. They suggest the possibility that experience with the tests in the first trial affects

[1] The calculation here described would in practice be a very cumbersome way of computing the square sum for the interactions ; it is much quicker to subtract the square-sums for the tests and persons at the end of the computation (see Mather, *loc. cit.*, p. 75 and Table 14A).

[2] I have discussed this particular interaction at some length because, as Mr. Mahmoud remarks, my previous example (this *Journal*, I, p. 23) has apparently led certain readers to suppose that only *one* elementary factor was needed to explain each interaction. In that earlier illustration no more happened to be needed, because, to simplify the exposition, I selected an inquiry in which only two tests were applied on each of two occasions ; hence all the interaction matrices there had a rank of one. For a detailed factor-analysis of a doubly-centred measurement matrix, such as that illustrated above, but obtained in a psychological research, see Burt, ' The Analysis of Temperament ', *Brit. J. Med. Psychol.*, XV, 1938, pp. 166 f., esp. Table I.

different children differently in the second. Hugh, for example, seems to profit by it; while James and George apparently find the second trial less interesting, and consequently show a decline. Here we are dealing with only two occasions; hence the interaction can be expressed by a single factor only.

The interaction between tests and occasions is quite devoid of significance. Had it been significant, we might tentatively have inferred that the abilities tested by the different subtests were affected differently by the passage of time: in other words, the results would have been suggestive of 'function fluctuation'. (Any decisive proof of such changes would need a more elaborate experimental design: see below.)

It will be instructive to compare the figures for the more detailed analysis (Table III) with those for the previous analysis (Table II). In the previous analysis the five subtests were pooled (this abolished Q_t and Q_{to}); and the difference between the means for the two trials was provisionally taken to be non-significant (hence Q_o and Q_{po} were combined to yield Q_r). And whereas that analysis was based on two sets of figures only for each person, namely, his average mark at either trial, the present analysis is based on five sets of figures for each trial—ten sets in all. This has the effect of multiplying the former square-sums by 5. The relations between the two are shown in Table VI.

TABLE VI. COMPARISON OF ANALYSES

Source of Variation		Square Sums First Analysis		Square Sums Second Analysis	
P.	Persons	(Q_p)	$4244{\cdot}2\times5 = 21220{\cdot}9$		21220·9
O.	Occasions	(Q_r)	$2138{\cdot}5\times5 = 10692{\cdot}4$	(Q_o)	3798·5
PO.	Interaction			(Q_{po})	6893·9
Sum		(Q_s)	$6382{\cdot}7\times5 = 31913{\cdot}3$	(Q_s)	31913·3

(a) *Marks Corrected for Difference between Means for Trials.* Since the difference between the means for the two trials has proved fully significant, we are now justified in regarding it as a source of constant error rather than of random error. It can be eliminated by converting the pooled marks into deviations about the mean for each year. The inter-class correlation remains as before; but the intra-class correlation now comes into closer agreement with it. Using the same formulae we have

$$r_{int} = \frac{1432{\cdot}7}{\frac{1}{2}(3566{\cdot}0+2057{\cdot}0)} = 0{\cdot}5096,$$

or

$$\frac{(21220{\cdot}9-6893{\cdot}9)\div5}{(21220{\cdot}9+6893{\cdot}9)\div5} = 0{\cdot}5096.$$

(b) *Marks Corrected for Differences in Test Scales.* The more detailed analysis has revealed peculiarities in the several subtests, resulting from differences either in difficulty or in severity of marking: these can be eliminated by converting the marks to deviation-form. There are, too, obvious differences in the range or spread of the marks, which must introduce an arbitrary weighting when the raw marks are summed as they stand: these can be abolished by reducing the deviations to standard measure.

B. STANDARDIZED MARKS

Correlations. To save space I shall not here reproduce the amended version of Table I which this standardization provides.[1] The changes, however, will inevitably modify the means or 'final marks' for the several boys. Hence the reliability coefficient must be calculated afresh. But there is yet another procedure for calculating such coefficients, which is highly instructive, and may therefore be briefly considered at this point.

In constructing a new battery, an experienced investigator will almost always calculate and examine, not only the correlation between the final marks, but also the correlations between the different subtests. If the marks have been reduced to standard measure, these can be obtained at once by simply computing the product-sums, and dividing by N. When the tests have been applied on two or more occasions, it is convenient to tabulate the coefficients so obtained in accordance

[1] Any reader who wishes to use this example for teaching purposes can obtain from the editor a copy of the detailed marks in standard measure and of the detailed intercorrelations.

with the scheme originally suggested for purposes of a group factor analysis,[1] arranging all the intercorrelations for the same day in square submatrices astride the leading diagonal. When there have been only two such applications, the entire correlation matrix will be quartered into four submatrices. The coefficients for the first day will then be placed in the N.W. quarter, since these will presumably have a group factor of their own ; those for the second day in the S.E. quarter, since they will contain a different group factor ; while the cross-correlations will be placed in the N.E. and S.W. quadrants, since they will involve neither group factor, but only the common factor (and perhaps specifics in the diagonal). The next step is to sum values in each of the four matrices ; and, if a group factor analysis is to be carried out, the subtotals for the rows and columns will be required as well. We can then calculate what Pearson would have called the ' bimultiple correlation ' between the two sets of tests, with or without differential weighting. The use of differential weights we can defer until we consider the application of factor analysis. Mr. Mahmoud gives an excellent illustration of the procedure in the paper that follows (cf. Table II, pp. 121 f.). To save space, therefore, I shall here omit the detailed table of correlations and give only the final results. As will be seen from Mr. Mahmoud's example, once the correlation table has been constructed, we have merely to take the sum of the figures in the N.E. (or S.W.) quarter for the numerator, and the mean of the sums in the N.W. and the S.E. quarters for the denominator. With the present data, if we take the geometric mean of the two last sums, we obtain the ordinary intercorrelation

$$r = \frac{9{\cdot}8041}{\sqrt{(22{\cdot}0912 \times 18{\cdot}9680)}} = 0{\cdot}4789 ;$$

if we take the arithmetic mean, we obtain the intra-class correlation

$$r_{int} = \frac{9{\cdot}8041}{\frac{1}{2}(22{\cdot}0912 + 18{\cdot}9680)} = 0{\cdot}4776.$$

The difference is now almost negligible.

Analysis of Variance. We shall obtain exactly the same figure for the intra-class correlation, if we carry out a detailed analysis of variance along the same lines as before. Our method of standardization has completely eliminated the variance between tests and between occasions, and therefore also the interaction between occasions and tests. Hence there now remains only one main source of variation (the differences between the mean marks obtained by the several persons—the result in which we are chiefly interested) and two possible types of interaction. The effects of these three components are shown in Table VII.

TABLE VII. ANALYSIS OF VARIANCE FOR STANDARDIZED MARKS

Source of Variation	Sum of Squares (Q)	Degrees of Freedom	Mean Square (V)	Variance Ratio	Probability
P. Persons	6·0667	5	1·2133	22·39	<0·001
PT. Persons and Tests	·7038	20	·0352	—	—
PO. Persons and Occasions	2·1451	5	·4290	7·91	<0·001
PTO. Residual	1·0844	20	·0542	—	—
S. Total	10·0000	50	(·2000)		

As before, both the differences between the means for the persons and the interaction between persons and occasions are fully significant, the former exhibiting much the larger amount of variance. But the interaction between persons and tests has become so small that its mean square is even less than the mean square for the residuals.

We can now calculate the intra-class correlation as before from the square-sums for persons and for the interaction between pupils and trials. We obtain

$$r_{int} = \frac{Q_p - Q_{po}}{Q_p + Q_{po}} = \frac{6{\cdot}0667 - 2{\cdot}1451}{6{\cdot}0667 + 2{\cdot}1451} = 0{\cdot}4776.$$

Since the degrees of freedom are now the same for both square-sums, the previous ' correction for bias ' (substituting mean squares for square-sums) leaves the value for the coefficient unaltered.

The Factorial Analysis of Variance. We have seen that a reliability coefficient is intended to indicate the ratio of the estimated variance of the ' true ' measurements to the actual variance of

[1] C. Burt, ' Factor Analysis by Submatrices ', *Journ. Psychol.*, VI (1938), p. 349, Table I. To avoid a technical term which is apt to alarm students who are unfamiliar with matrix algebra, each ' submatrix ' whose elements were to be summed was called a ' pooling square '. The phrase seems now to be more frequently used of the entire correlation matrix as thus partitioned.

the observed measurements, i.e., to the ' total variance ' conceived as the sum of the ' true variance ' and an ' error variance '. But how do we know that the value taken as the numerator in the ratio just calculated really represents the ' true variance ' we have in mind, and that it does not incorporate something that we might (if we knew its real nature) also regard as error?

It is, I think, seldom recognized that an ' analysis of variance ', as ordinarily carried out, does not really indicate the relative amounts of the total variance attributable to each hypothetical factor taken by itself. Its purpose is primarily to test the statistical significance of such factors, not to assess their relative importance. For this latter purpose the ' square sums ' and the ' mean squares ' must be subdivided still further.

The need for a supplementary analysis will be clearer if we first set out, in the form of a hypothetical model, our conception of the way the observed measurements are composed. Adopting the same style of notation as before, our hypothesis will be that x_{ijk}, the observed measurement for the ith person obtained with the jth test on the kth occasion, is made up of four additive factors, each of which is normally distributed in the population with a mean of zero, and varies independently of variations in the others. Thus, we may write

$$x_{ijk} = \gamma_i + \tau_{ij} + \omega_{ij} + \epsilon_{ijk},$$

where the new symbols designate the following factors :

(i) A general or basic factor, γ_i, specifying the extent to which the true mean measurement for the ith individual deviates from the mean of the population, ξ (which we may, if we wish, set equal to zero :) i.e.,

$$\gamma_i = \xi_i - \xi. \qquad \text{(viii)}$$

(ii) A supplementary factor, τ_{ij}, possibly complex, specifying the extent to which the ith individual's measurement for the jth type of test deviates from what we should expect in terms of his own general mean and of the population mean for that test (ξ_j) : i.e.,

$$\tau_{ij} = \xi_{ij} - \xi_i - \xi_j + \xi. \qquad \text{(ix)}$$

(iii) A supplementary factor, ω_{ij}, also possibly complex, specifying the extent to which the ith individual, tested with all the tests on the kth occasion, would deviate from what we should expect in terms of his own general mean and of the population mean for that occasion (ξ_k) : i.e.,

$$\omega_{ik} = \xi_{ik} - \xi_i - \xi_k + \xi. \qquad \text{(x)}$$

(iv) An error factor, ϵ_{ijk}, specifying the extent to which the observed measurement for the ith individual, obtained with the jth test on the kth occasion, deviates from what we should expect in terms of the three foregoing factors : i.e.,

$$\epsilon_{ijk} = x_{ijk} - (\gamma_i + \tau_{ij} + \omega_{ik}). \qquad \text{(xi)}$$

From this it will be clear that the observed means and variances cannot, as they stand, be taken as unbiased estimators for the corresponding factor measurements, since each of them includes a contribution from the other factors, usually in diminishing amount. Thus

$$\bar{x}_i = \frac{1}{mn} \sum_j^n \sum_k^m x_{ijk},$$

$$= \gamma_i + \frac{1}{n} \sum_j^n \tau_{ij} + \frac{1}{m} \sum_k^m \omega_{ik} + \sum_j^n \sum_k^m \epsilon_{ijk}. \qquad \text{(xii)}$$

Squaring and summing the several means and their analytic expansions in terms of these factors, we readily obtain the requisite equations for estimating the variances of the several factors. Then, substituting the values given for the variances in Table VII, we reach the estimates shown below (Table VIII), where, as before, $V_p = Q_p/(N-1)$, $V_{pt} = Q_{pt}/(N-1)(m-1)$, $V_{po} = Q_{po}/(N-1)(m-1)$, $V_{pto} = Q_{pto}/(N-1)(m-1),(n-1)$, and hence $V_s = Q_s/(N-1)mn$. As a check, we note that the value now obtained for the sum of the several variances should agree with the value for V_s given in Table VII.

I have deliberately given the results of a mechanical application of the formulae to the present data, because it illustrates one or two instructive points. First we observe that an impossible estimate is reached for s_τ^2. The negative value is due to the fact that we have treated the results of a chance fluctuation as representing a separate factor : it implies that, with the data here analysed, it was unnecessary to include in the hypothetical model any such factor as τ_{ij}. This, of course,

should have been inferred from the attempt to determine the significance of the interaction between persons and tests in Table VII. But, since several correspondents have been puzzled by the occasional appearance of a negative result at this stage, it seemed useful to illustrate how it may arise.

TABLE VIII. ESTIMATES FOR THE VARIANCES OF THE POSTULATED FACTORS

Variance	Equation	Estimate
s_γ^2	$= \frac{1}{mn}(V_p - V_{pt} - V_{po} + V_{pto})$	·0803
s_τ^2	$= \frac{1}{m}(V_{pt} - V_{pto})$	−·0095
s_ω^2	$= \frac{1}{n}(V_{po} - V_{pto})$	·0750
s_ε^2	$= V_{pto}$	·0542
$\sum s^2$	$= V_s$	·2000

Secondly, let us apply the foregoing analysis to the interpretation of the ordinary reliability coefficient. We now have [1]

$$r_{xx'} = \frac{V_p - V_{po}}{V_p + (m-1)V_{po}} = \frac{ns_\gamma^2 + s_\tau^2}{n(s_\gamma^2 + s_\omega^2) + s_\tau^2 + s_\varepsilon^2} \quad . \quad . \quad . \quad . \quad \text{(xiii)}$$

$$= \frac{\cdot 4051 - \cdot 0095}{\cdot 4015 + \cdot 3750 - \cdot 0095 + \cdot 0542}$$

$$= \cdot 4776 \text{ as before.}$$

An equation like the above, expressed in terms of what may be called factorial variances, yields, I venture to think, a far clearer conception of what such coefficients are measuring than the more familiar equations which relate them to the cruder results furnished by an ordinary analysis of variance.[2] The detailed figures here obtained, for instance, bring into sharp relief the question raised above. When we speak of the ' reliability ' of a test, we imply that it is a more or less ' reliable ' measure of something ; but of what? Are we seeking a reliable measure merely of the basic factor (γ) common to all the tests? Or are we intending to secure a more general estimate, which would include the specialized abilities of the separate tests as well?[3] In other words, is τ (if significant) part of the ' true measurement ', or is it irrelevant and therefore part of the ' error '? In the former case the figure obtained for the reliability would here be too low ; in most cases it would be too high.

Reliability determined from a Single Application. There have been many attempts to derive an estimate of reliability, or at least of its upper or lower limits, from the results obtained by administering a composite test (or battery) once only. I have discussed this problem more fully elsewhere,[4] and therefore need say very little here.

The relations between this procedure and the older and more familiar method may be most conveniently explained in terms of the pooling square. At first sight the idea that the consistency of two or more test-applications can be estimated from a single application only seems highly paradoxical.

[1] Since this section is intended to be of general application, I here give the general form of the equation for the unbiased intra-class correlation (see *Factors of the Mind*, p. 275, eqn. ii). Since in the present example $m = 2$, the equation reduces to the forms already used above (pp. 109 and 113).

[2] The relations between the two types of component are very similar to the relations between the summational or centroid factors obtained by ' simple summation ' and the group factors obtained by a ' group factor analysis ' : with the former the ' general factor ' includes not only the ' basic ' factor but also contributions from the group factors, and similarly with the successive bipolars. This is confirmed by the results of an actual factor analysis by these two methods : see below, p. 118.

[3] For example, when using a battery of verbal and non-verbal intelligence-tests, our aim is generally to measure intelligence only ; when using question-papers in arithmetic, English, and the like to measure ' general educational ability ' in a scholarship examination, our aim generally is to assess a compound of all the child's educational attainments. Elsewhere I have tried to distinguish the various types of so called ' reliability coefficient ', and to give them more distinctive names : (see also Mr. Mahmoud's paper below, pp. 127f.).

[4] See Burt, 'The Reliability of Teachers' Assessments', *Brit. J. Educ. Psychol.*, XV (1945), esp. *Appendix*, pp. 90–2.

It will be remembered, however, that our definition treats reliability as essentially a characteristic of the test, and accordingly, in estimating its value from two applications, we have assumed that the persons do not change significantly in the interval and that the test or subtests applied on the second occasion do not differ significantly from those applied on the first, or, if there *are* any changes in the persons or any differences in the tests, then they must be eliminated (so far as possible) by an appropriate statistical formula. Now, if there were no changes and no differences *whatever*, then the correlations between the subtests on the second occasion would be identical with those obtained on the first, and the cross-correlations between two applications of the same test would take the form of 'reduced self correlations', i.e., their amount would be determined solely by the factors common to all the subtests, and would be unaffected by factors specific to either the first trial or the second.

With these assumptions we can now construct a suitable pooling square. The N.W. quadrant will, as before, contain the observed inter-correlations obtained from the first application, which is now the only *actual* application : as usual, it will have units (1·00) in the leading diagonal. The S.E. quadrant will represent the results of an *imaginary* second application carried out with perfectly equivalent subtests and under perfectly identical conditions. Hence all the correlations in the S.E. quadrant will be exactly the same as in the N.W. quadrant, and the side correlations in the N.E. quadrant will be exactly the same as the side correlations in both the N.W. and the S.E. quadrants : in the N.E. quadrant, however, the leading diagonal will no longer contain units, but 'reduced self correlations' conforming with the assumptions just laid down. The only question is—what values are we to substitute?

If we already possess estimates for the reliability of the separate subtests, we might perhaps use these ; and their combined unreliability we may reasonably equate with s_ε^2. In this case the numerator for the reliability-ratio will differ from the denominator only by the omission of the term s_ε^2. We shall have in fact

$$r_{xx'} = \frac{\sum R_{ab}}{\sum R_{aa}} = \frac{n\,(s_\gamma^2+s_\omega^2)+s_\tau^2}{n\,(s_\gamma^2+s_\omega^2)+s_\tau^2+s_\varepsilon^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xiv)}$$

Evidently this procedure will yield a higher value for the reliability than that obtained from two *actual* applications : for in the numerator of equation (xiii) (p. 116) the term ns_ω^2 was also omitted : when the requisite conditions are fulfilled, therefore, we may regard eqn. xiv as indicating an upper boundary.

However, as in factor analysis, it would seem more in accordance with general theory to eliminate from the diagonal elements not only the unreliability of each subtest but also its 'specific factor' (i.e., the group factor common to the two trials). Indeed, this is all we can hope to attempt if we possess no previous estimates for the unreliability, since in that case s_ε^2 will be merged with s_τ^2. We must therefore reduce the numerator of our equation to $n(s_\gamma^2+s_\omega^2)$. To calculate this value, in the absence of any other empirical guide, the simplest practical device will be to take the average of the side-correlations as providing the best estimate for the average of the reduced self correlations. In terms of the hypothetical factor-variances, the formula for the reliability coefficient will then be

$$r_{xx'} = \frac{n\,(s_\gamma^2+s_\omega^2)}{n\,(s_\gamma^2+s_\omega^2)+s_\tau^2+s_\varepsilon^2}. \quad . \quad . \quad . \quad . \quad . \quad \text{(xv)}$$

Since, in the absence of all evidence to indicate their relative proportions, the distinction between s_γ^2 and s_ω^2 and between s_τ^2 and s_ε^2 now becomes meaningless, we may re-write the formula in the simplified form [1]

$$r_{xx'} = \frac{n\,s_G^2}{ns_G^2+s_E^2}. \quad . \quad . \quad . \quad . \quad . \quad \text{(xvi)}$$

It is tempting to regard this as indicating the lower boundary. It must be remembered, however, that the factor G will now very probably incorporate most of the former factor ω and possibly much of the former factor τ. Hence, the estimated reliability may still be magnified to an unknown extent by the inclusion in the 'true' common-factor variance of contributions due to the special conditions obtaining at the time of the single trial. This obvious inference appears to have been ignored by many who advocate the simplified procedure.[2]

[1] This is the equation reached in the *Appendix* just cited (p. 90, eqn. v). The effect of using the average intercorrelation as described above is there indicated in eqn. viii ; and the worked example given in the body of the article illustrates the relation of the coefficient so obtained to the results of the analysis of variance.

[2] Spearman's 'split-half' method is a special instance of this method, and is open to the foregoing (as well as other) objections. As I have shown elsewhere, the procedure described in the text and the formulae so deduced give the average value of the coefficients obtained *by all possible splits*.

The point is clearly illustrated by the present data, where the two trials were separated by an interval of 12 months (not, as our assumptions require, by the minimum time needed to prevent the effects of the first trial influencing the second). The sum of the 10 inter-correlations (without the self-correlations) obtained at the first trial was 17·09. Accordingly with the first method (inserting actual reliabilities for subtests) we obtain

$$\frac{17{\cdot}09+1{\cdot}77}{17{\cdot}09+5\times 1{\cdot}00}=\frac{18{\cdot}86}{22{\cdot}09}=\cdot 854\,;$$

with the second (inserting estimates based on the average intercorrelation)

$$\frac{17{\cdot}09+5\times{\cdot}85}{17{\cdot}09+5\times 1{\cdot}00}=\frac{21{\cdot}34}{22{\cdot}09}=\cdot 966.$$

Reliability determined from Weighted Factors. The factor measurements obtained or implied by an analysis of variance consist of the simple sums or means of the observed test-measurements, without any differential weighting. As I have argued elsewhere, better estimates of reliability can be secured if we employ weighted sums, the appropriate weights being determined by factor analysis. Here, for example, the summational and the group factor analysis alike reveal that, on both occasions, Test *P* has a far higher loading for the general or basic factor than Test *T*; and this in turn greatly augments the size of the interaction between tests and persons.

It is largely for this reason that I have repeatedly urged the need, in any exact inquiry, to base the estimates of reliability on an explicit factorial analysis by a summational or group factor method.[1] But other questions frequently arise which only a factor analysis can solve. In an ordinary analysis of variance the most suitable form of classification is often a matter of some doubt—a point that is commonly overlooked. The 'degrees of freedom' can always be subdivided into 'orthogonal comparisons' in many different ways: and, as Mather observes, "the first task is to partition these degrees of freedom into the components that are appropriate to the various comparisons which might be interesting".[2] But how can we tell which comparisons "might be interesting"? Shall we examine those that represent the special abilities conceivably required for particular groups of tests' and if so which groups do we select? Shall we look for those that represent 'function-fluctuation' in the persons, or, it may be, fluctuations in the conditions obtaining at successive trials, which themselves may affect different persons differently? Or again how are we to analyse the interactions into their constituent components and assess the importance of each? With large groups, or numerous subtests, or numerous repetitions, such problems often prove crucial. And nothing but a preliminary factor analysis can show how they are to be answered or which of the supplementary classifications are really worth while.

Since our initial table of data was based on so small a sample and does not represent any actual testing, I need not here reproduce the tables of correlations and factor saturations. Moreover, the method has been fully illustrated in previous papers.[3] For completeness I will merely summarize the final results. Three summational factors and three group factors were extracted. Their contribution to the total variance (10·00) was as follows:

(A) General factor, 6·11; first bipolar (classifying occasions), 2·37; second bipolar (classifying tests), 0·35 (non-significant);

(B) Basic factor, 4·48; group factor for the first occasion, 1·88; group factor for the second occasion, 2·33.

It will be seen that the variances of the general factor and the bipolar factor for occasions stand in approximately the same proportions as the crude variances V_p and V_{po} in Table VII; while the basic factor and the two group-factors for occasions stand in almost exactly the same proportions as s_γ^2 and s_ω^2. This is what, if the foregoing theory is correct, we should expect. So far as it goes, these points of agreement confirm the close relation between the principles underlying factor analysis and those underlying the analysis of variance. However, in many actual researches, where the diagnostic values and consequently the weights of the several subtests differ widely from one to another the agreement no longer obtains; and in such cases a reliability coefficient based on a factorial analysis will manifestly be far more accurate and far more informative than one derived by the more usual over-simplified procedures.

[1] Which factorial method is to be preferred will depend on the investigator's conception of the factor or factor measurements whose reliability he desires to assess. Thus I should not entirely agree with Mr. Mahmoud who appears to hold that the group factor method alone is appropriate for *all* such problems.

[2] *Op. cit.*, p. 73. Kendall observes that the choice is 'as a rule determined by some prior classification given among the data of the problem' (*Advanced Theory of Statistics*, II, 175). But in psychology, instead of being given *with* the data, the classification has often to be deduced *from* the data; and for this purpose a factorial analysis may be indispensable.

[3] An excellent illustration is provided in this issue by the tables included in Mr. Mahmoud's article. If any reader wishes to use the correlational tables relating to the present data for purposes of class-instruction, copies can be obtained on request.

Vol. VIII Part II The British Journal of Statistical Psychology November 1955

TEST RELIABILITY IN TERMS OF FACTOR THEORY

By A. F. MAHMOUD

The Need for a More Adequate Theory. In a previous issue of this *Journal*[1] I raised the question of assessing the reliability of composite tests (like the newly constructed battery of intelligence tests on which my colleagues and I have been engaged); and the Editor in reply drew attention to the possibility of using more refined techniques, such as have recently been developed from the familiar principles of factor analysis and analysis of variance. In several published researches one or other of these newer methods has been tentatively adopted. But the investigators have usually failed to extract all the information implicit in their data, either because the methods have not been carried far enough, or because they have been applied too mechanically and without an adequate recognition of the assumptions or hypotheses on which the procedures are based. It is the object of this note to restate the underlying theory[2] a little more simply, and then to show, by an actual example, first the value of the supplementary device that Professor Burt has termed the 'factorial analysis of variance', and secondly the superiority, for this particular problem, of a 'group factor analysis' rather than a summational or centroid procedure.

Illustrative Example. For the purposes of the inquiry referred to in my note, eight intelligence tests were constructed, and arranged to form two parallel batteries with similar subtests in each. Each of the subtests consisted of 20 problems or tasks of increasing difficulty. As they were intended for a comparatively illiterate population, all were of a non-verbal or performance type. Test 1 consisted of a picture-completion test—four stories each containing five panels in series with an inset to be chosen and inserted in each; Test 2 of geometrical designs to be constructed with coloured blocks, similar to Kohs'; Test 3 of 20 mazes, similar to the Porteus series; Test 4 of diagrammatic 'matrices' similar to Raven's series. A hundred boys were selected as forming a random sample of the population, and were tested with the first battery as nearly as possible on their eleventh birthday and again with the second battery some six months later. Thirteen were absent on the second occasion, so that the data analysed here relates to 87 individuals only. In scoring the test, arbitrary marks of 2 to 5 were awarded for each task correctly performed, the amount for each being determined partly by its complexity and partly by the testee's speed. Taking the unweighted sum of each pupil's marks for each subtest as his 'final mark' for the battery, and accepting the teachers' assessments for intelligence as a provisional criterion, we obtained validity coefficients of 0·56 for the first trial and 0·63 for the second—an average of 0·59. This is a somewhat disappointing figure. But as we shall see later on, better correlations are obtained when the subtests are given different weights; and, since the children were drawn from different schools and the teachers' standards varied appreciably, the criterion does not provide very trustworthy estimates of individual differences, and can only be regarded as furnishing a convenient overall check.

(a) *Analysis of Variance.* As a preliminary it seemed advisable to determine (i) whether the subtests comprising the two batteries showed any appreciable differences either in the general difficulty of the tasks or in the general method of marking, and (ii) whether the abilities that we sought to test showed any appreciable change with time.[3] We therefore began by making a complete three-way

[1] 'A Reliability Coefficient based on Analysis of Variance', *Brit. J. Statist. Psychol.*, VII (1954), p. 61.

[2] I am much indebted to Professor Burt for supplying copies of his original memorandum on Reliability of Tests and Examinations, drawn up for the International Institute Examinations Enquiry, and of his later Laboratory Notes, and for permitting me to incorporate in my own presentation whatever extracts I wished. His abridged 'three-way analysis of variance' and what he calls 'the factorial analysis of variance' appear to be especially appropriate to problems like my own; and, save for a few unpublished theses and reports, do not seem to have been used in any large scale research.

In adapting his arguments, a few modifications have been made, chiefly in the interests of simplification: his notation, for example, distinguishes between (*a*) constants (or 'parameters') which characterize the population, whether actual or hypothetical, and (*b*) the corresponding estimates (or 'statistics') which are obtained from the sample. Here, to avoid complicating the discussion too much for the elementary reader, I have abandoned this double symbolism. The statistical reader will easily be able to make such distinctions for himself where he considers them important.

[3] As will be obvious later on, it is scarcely possible to distinguish between the two explanations for the change in the averages, unless a more complicated experimental design is used. For such a purpose we should need at least four groups of children—one pair taking one battery first, the other pair taking the other battery first.

analysis of the variance for the raw scores. The working method followed was that recommended by Professor Burt in his Notes on *The Construction and Standardization of Tests* (illustrated in the preceding paper).

In analysing such data there are three main sources of variation to be considered—the persons tested (P), the tests employed (T), and the occasions on which the tests were administered (O). The crude analysis, based on the raw unstandardized scores, indicated (i) that the differences between the means for the several persons were highly significant ; (ii) that the differences between the means for the several subtests were non-significant ; (iii) that the differences between the means obtained on the two successive occasions were barely significant at the 5 per cent. level ($P = 0{\cdot}047$) : there was in fact a slight increase in the general average, due, it appeared, not so much to mental development or to previous practice as to slight differences in the standardization.

Accordingly, it seemed reasonable to eliminate these minor differences between tests and occasions as irrelevant, and reduce all scores to standard measure. In computing the standard deviations from the raw scores, the observed sums of squares were divided by $(N - 1)$, not by N. The square-sum for each subtest will therefore be 86 and the mean zero. A fresh analysis of variance was then carried out. The results are shown in Table 1.

Since we have eliminated all differences between the means for subtests and for trials, and consequently the interaction between subtests and trials, we now have to deal with only four sources of variation, and the total number of degrees of freedom is reduced to

$$(N-1)mn = (87-1)\times 4\times 2 = 688.$$

With the degrees of freedom indicated for the separate sources of variation, all three variance ratios appear to be fully significant even at the 0·1 per cent. level.

TABLE I. ANALYSIS OF VARIANCE

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	Variance Ratio
Persons (P)	480·069	86	5·5822	46·7
Persons and Subtests (PT)	129·748	258	·5029	4·2
Persons and Trials (PO)	47·326	86	·5503	4·6
Residual (PTO)	30·857	258	·1196	—
Total	688·000	688	(1·0000)	—

Note.—The value for each 'Mean Square' (in column 3) is obtained by dividing the 'Sum of Squares' by the 'Degrees of Freedom' : hence the figure in the last line is not a 'Total' of those above.

Using the formula given by Burt,[1] we can calculate the correlation between the two batteries (i.e., the ordinary reliability coefficient) from the variances in the table, without having to compute the product-sums. We obtain

$$r_{tt'} = \frac{V_p - V_{po}}{V_p + V_{po}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(i)}$$

$$= \frac{5{\cdot}5822 - {\cdot}5503}{5{\cdot}5822 + {\cdot}5503} = {\cdot}8205, \text{ where}$$

V_p and V_{po} denote the mean squares for persons and for the interaction between persons and trials.

The figure thus obtained is no doubt quite promising for a first attempt with a new set of tests ; but (assuming we can accept it as it stands) it is too low for us to consider the battery sufficiently reliable for immediate use. Moreover, the analysis of variance reveals a feature which the reliability coefficient ignores—namely, the large interaction between persons and tests. This certainly calls for further scrutiny. Does it represent one additional and unsuspected factor or several? And does its magnitude fairly reflect that of the disturbing influences to which it points?

Psychologists who have used analysis of variance seem generally to suppose that the variances thus assigned to the several 'sources of variation' can always be identified, just as they are, with the variances of the hypothetical components or 'factors' that we are seeking to determine.[2] As a

[1] *Factors of the Mind*, p. 275, remembering that k (as there defined, our m) = 2. Cf. *Marks of Examiners*, p. 255. [See also article, p. 109].

[2] Burt's comparative tables for 'Analysis of Variance with Standardized Marks' and 'Analysis of Variance with Factors obtained by Weighted Summation' (this *Journal*, I, 12 and 25, Tables VI and XV) have perhaps encouraged this view. With the data there tabulated the interaction between persons and tests and between persons and occasions could be expressed in terms of a single row vector, analogous to a row of 'factor measurements' for a single factor in either case, and the variances were directly compared with the variances obtained from a factor analysis of the usual kind. But this was no doubt due to the need for a short and simplified exposition. In his fuller analysis of the Examinations Council data he stresses the need for what he calls the 'further analysis of variance' to obtain the purified 'factorial variances'.

matter of fact, however, when the analysis stops at the usual point, the ' square sums ' calculated for the larger sources of variation always incorporate contributions from the smaller sources (i.e., from the interactions and residuals), much as the larger factors that are first extracted with a summational or centroid method incorporate contributions from the lesser and later group factors. Hence, to estimate the true magnitude of the hypothetical components the analysis must be carried one stage further. The importance of this for the assessment of reliability will be clearer if we turn for a moment to the more familiar approach—the examination of the relevant correlations.

(b) *Correlational Analysis.* Few investigators, in constructing a new battery, attempt a formal analysis of variance ; but nearly all proceed to calculate the correlations between the various subtests. When the battery has been administered on two or more occasions, the coefficients may conveniently be arranged according to the scheme adopted by Burt in carrying out a ' group factor analysis '. The figures obtained in the present research are set out in Table II. The correlation matrix, it will be seen, has been partitioned into four submatrices or ' quadrants '. The intercorrelations for the first trial are entered in the N.W. quadrant, those for the second trial in the S.E. quadrant, and the cross-correlations in the N.E. and S.W. quadrants.

TABLE II. CORRELATIONS BETWEEN TESTS : (POOLING SQUARE)

Tests	1	2	3	4	Total (1st Trial)	1	2	3	4	Total (2nd (Trial)
Test 1. Picture	[1·000]	·897	·625	·713	3·235	·881	·768	·604	·637	2·890
2. Block	·897	[1·000]	·541	·584	3·022	·825	·837	·539	·563	2·764
3. Maze	·625	·541	[1·000]	·751	2·917	·445	·347	·703	·557	2·052
4. Matrix	·713	·584	·751	[1·000]	3·048	·559	·462	·667	·670	2·358
Total (1st Trial)	3·235	3·022	2·917	3·048	12·222	2·710	2·414	2·513	2·427	10·064
1. Picture	·881	·825	·445	·559	2·710	[1·000]	·774	·641	·735	3·150
2. Block	·768	·837	·347	·462	2·414	·774	[1·000]	·572	·704	3·050
3. Maze	·604	·539	·703	·667	2·513	·641	·572	[1·000]	·728	2·941
4. Matrix	·637	·563	·557	·670	2·427	·735	·704	·728	[1·000]	3·167
Total (2nd Trial)	2·890	2·764	2·052	2·358	10·064	3·150	3·050	2·941	3·167	12·308

The coefficients in each quadrant have been summed, in order to show how the sums of the correlations between the subtests may be used to obtain the correlation between the two entire tests. The ordinary proof of the formula for this purpose, as given by Pearson, Kelley and others, is lengthy and difficult for the ordinary student to follow.[1] It may, however, readily be simplified. The final marks which we desire to correlate are simply the unweighted [2] sums or averages of the marks obtained by each boy in the four subtests on the first and second occasions respectively. Accordingly, let w' (a row vector) denote the summation operation [1, 1,..., 1], let M_a (an $n \times N$ matrix) denote the matrix of marks obtained with the first set of tests applied on the first day, and M_b that obtained with the second set applied on the second day. Then

$$r_{ss'} = \frac{w' M_a M_{ab} w}{\sqrt{(w' M_a M_a' w \cdot w' M_b M_b' w)}} = \frac{w' R_{ab} w}{\sqrt{(w' R_{aa} w)(w' R_{bb} w)}} \quad . \quad . \quad . \quad \text{(ii)}$$

[1] Equation (ii) is merely the well-known formula for the correlation between sums (T. L. Kelley, *Statistical Method*, 1923, p. 197, eqn. 1.47). Burt gives the proof quoted in the text to show how the adoption of matrix notation simplifies the older algebraic demonstration ; here, it will be seen, a derivation, which with Kelley occupies nearly two pages, is condensed into a couple of lines.

[2] When weights are used to maximize this correlation, the result is the ' canonical correlation ' between the two batteries : cf. C. Burt, ' Factor Analysis and Canonical Correlations ', this *Journal*, II (1948), 95-106 : the formula for the unweighted correlation, derived by this method, is given on p. 97, eqn. iii.

or, to adopt the notation of the articles just cited,

$$r_{ss'} = \frac{\sum R_{ab}}{\sqrt{(\sum R_{aa} \sum R_{bb})}}. \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(ii } a\text{)}$$

Thus, with the present example, the square sums for the two batteries are given by the sums of the intercorrelations shown in the N.W. and S.E. quadrants respectively of Table II ; while the product sum is given by the sum of the cross correlations in the N.E. (or S.W.) quadrant.

Hence the correlation between the two batteries will be

$$r_{ss'} = \frac{10{\cdot}064}{\sqrt{(12{\cdot}222 \times 12{\cdot}308)}} = {\cdot}8205.$$

It will be noted that this is exactly the same value as was obtained from the analysis of variance by using equation (i). An algebraic proof of their equivalence will be given in a moment. Our more immediate object is to factorize this matrix of correlations in terms of the hypothetical components corresponding to those sources of variation which the preceding analysis of variance has shown to be significant.

Correlation as a Ratio of Variances. Fisher,[1] when first introducing his method of analysing variance, begins by explaining how a coefficient of correlation can be expressed as a ratio of two variances, or in other words, as " that fraction of the total variance $(A+B)$ which is due to the cause (A) which the observations have in common" : i.e., in symbolic form, "$\rho = \frac{A}{A+B}$". He then points out that, whereas " the value of B (the ' error variance ' so-called) may be estimated directly from the variance *within* groups ", the value found for A, the variance *between* groups, includes a part contributed by B. The argument that follows may be regarded as a generalization of this analysis to the case in which we are concerned with three sources of variation (in addition to 'error' variance).

Hypothetical Model. In carrying out such an analysis, and particularly in applying it to the determination of reliability, the most urgent need, in my opinion, is to keep clearly in mind the underlying assumptions on which the procedure is based and to consider how far they hold good for the data under examination. Strangely enough, these obvious precautions seem rarely to be observed. Let us therefore start by attempting an explicit formulation of the assumptions involved.

For our present purpose, we may most appropriately begin with the simplest of Burt's ' mathematical models ' (as he terms them). It is based on the provisional hypothesis that any observed test-measurement, obtained in the course of an investigation like the present, can be regarded as the weighted sum of four kinds [2] of hypothetical component—(i) a factor or factors [3] common to all the tests on all the occasions ; (ii) a group factor characteristic of similar specimens of the same test (with n kinds of test there may be n such components) ; (iii) a group factor representing those conditions that may be peculiar to the same occasion : (when the tests are applied on two or more occasions there may be two or more such components) ; (iv) a specific factor peculiar to each single test ' error of measurement ' : (with n tests and 2 occasions there will be $2n$ such components).

Let us then suppose that we are dealing with a battery of n tests, and that these are applied in identical or parallel form on two (or more) different days ; and let m_{ijk} denote the observed mark or measurement [4] obtained by the ith pupil in the jth test on the kth day ($i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, m$). Let x denote the hypothetical factor measurements ; and let g, d, t and e denote the relative weights (or factor saturations) of the four kinds of factor. Let the subscript i denote the individual to whom the measurements refer, and the subscript j or j' the test to which the saturations refer, the subscript k being added where it is necessary to distinguish days. Then, for any given day, we may write

$$m_{ij} = g_j x_{gi} + t_j x_{ti} + d_j x_{di} + e_j x_{ji} \quad . \quad . \quad . \quad . \quad . \quad \text{(iii)}$$

[1] *Statistical Methods for Research Workers*, chap. VII, sect. 40.

[2] This is merely a special case of the ' four factor theory ' set out by Burt in his discussion of reliability in *Marks of Examiners*, p. 259, eqn. 5.

[3] In the algebraic equations to be developed from these assumptions I shall for simplicity assume that there is only *one* component common to all the tests, or that, if there are more, their variances can be summed and treated as the variance of a single composite component. This simplification is warranted by Assumption 1. However, it means that, if there are factors common to two or more tests in the same battery but not to all, our present model must treat them as bipolar factors extending over all the tests (as is done in an ordinary centroid analysis). We could, of course, avoid this by introducing (as Burt does in one of his more elaborate models) a fifth kind of factor, namely, ' group factors for special abilities '. But I am anxious not to obscure the exposition of the main argument by raising numerous minor issues.

[4] It seems a little inelegant, but not, I think, confusing, to use m (with subscripts) for ' marks ' or ' measurements ' and m (without subscripts) for number of occasions or trials. These symbols have been adopted in most of the papers, etc., to which I shall here refer. Hence it seemed best to retain them.

Following the usual convention, let us suppose that both the test-measurements and the hypothetical factor-measurements have been reduced to unitary standard measure, so that for any given day $\sum_i m_{ij}^2 = 1$, $\sum_i x_{gi}^2 = 1$, etc.

Assumption 1. In dealing with observable classifications (such as are used in analysis of variance) it may often happen that the crude *empirical* factors (i.e., the principles of classification) are not altogether independent of each other. For purposes of analysis, however, we shall assume that *the hypothetical components are strictly independent, i.e., mutually uncorrelated.*[1] Then all such product sums as $\sum_i x_{gi}x_{di}$, $\sum_i x_{ti}x_{ji}$, $\sum x_{ji}x_{j'i}$, etc., will vanish. Hence, for any given day,

$$\sigma_j^2 = \sum_i m_{ii}^2 = g_j^2+t_j^2+d_j^2+e_j^2 \quad . \quad . \quad . \quad . \quad . \quad \text{(iv)}$$

i.e., the variance of the jth test will be the sum of the variances of its four components. And similarly the correlation between two different tests applied on the same day will be

$$r_{jj'} = \sum_i m_{ij}m_{ij'} = g_jg_j+d_jd_j. \quad . \quad . \quad . \quad . \quad . \quad \text{(v)}$$

between two similar tests applied on different days will be

$$r_{j_1j} = \sum_i m_{ij_1}m_{2ij} = g_{j_1}g_{i_2}+t_{j_1}t_{j_2}. \quad . \quad . \quad . \quad . \quad \text{(vi)}$$

And between two different tests applied on two different days it will be simply

$$r_{j_1j'_2} = \sum_i m_{ij_1}m_{ij'_2} = g_{j_1}g_{j'_2}. \quad . \quad . \quad . \quad . \quad . \quad \text{(vii)}$$

The model we have adopted, together with this first assumption, will enable us to express the sums derived from the four quadrants of our ' pooling square ' (cf. Table II above) in terms of four kinds of hypothetical component. Summing the n values for σ_j^2 and $r_{jj'}$, and substituting the expressions given in eqns. (iv) to (vii), we obtain the results shown below in Table III (reproduced from Burt). It will be observed that, in virtue of the derivation of the factorial expansions for the various coefficients, the expressions $\sum t_{j_1}^2$, $\sum e_{j_1}^2$, $\sum t_{j_2}$, and $\sum e_{j_2}$ enter only into the diagonals of the N.W. and S.E. quadrants respectively, and the expression $\sum t_{j_1}t_{j'_2}$ enters only into the diagonal of the N.E. and S.W. quadrants : in other words, these values are regarded as ' specifics ' which merely augment the ' self-correlations ' in the matrices R_{aa}, R_{ab}, and R_{bb}. To determine the best values for each of the several factor saturations, a formal factor analysis by the group procedure will be essential (cf. Table VI below, p. 133). Meanwhile, if we introduce two further simplifications, we can reach a plausible approximation by means of a direct analysis of variance.

TABLE III. POOLING SQUARE WITH DIFFERENTIAL SATURATIONS

Sums of Correlations in terms of Hypothetical Saturations

Factor Saturations	Summed Correlations	
$g_{11}\ d_{11}\ -\ t_{11}\ -\ e_{11}\ -\ -\ -$ $g_{21}\ d_{21}\ -\ -\ t_{21}\ -\ e_{21}\ -\ -$	$\sum R_{aa} =$ $(\sum_j g_{j_1})^2+(\sum_j d_{j_1})^2+\sum_j t_{j_1}^2+\sum_j e_{j_1}$	$\sum R_{ab} =$ $\sum_j g_{j_1}g_{j_2}+\sum_j t_{j_1}t_{j_2}$
$g_{12}\ -\ d_{12}\ t_{12}\ -\ -\ -\ e_{12}\ -$ $g_{22}\ -\ d_{22}\ -\ t_{22}\ -\ -\ -\ e_{22}$	$\sum R_{ba} =$ $\sum_j g_{j_1}g_{j_2}+\sum_j t_{j_1}t_{j_2}$	$\sum R_{bb} =$ $(\sum_j g_{j_2})^2+(\sum_j d_{j_2})^2+\sum_j t_{j_2}^2+\sum_j e_{j_2}^2$

[1] For a rigorous treatment it would be advisable to make it perfectly clear whether such conditions as this refer to the population or to the sample. That, in fact, is the procedure followed by Burt. But, as mentioned above, such a distinction entails a double set of symbols (Greek for the former, Roman for the latter) which is likely to bewilder the ordinary reader. It is perhaps sufficient to note that the *model* itself refers primarily to the population. Largely for convenience in subsequent deductions, certain modes of factor analysis also assume that the hypothetical factor measurements are uncorrelated in the sample. When the model is applied in any particular research, the investigator should first satisfy himself that the conditions postulated hold good within a reasonable approximation. For purposes of simplified calculation the requisite assumptions can then be supposed to be true also of the data contained in the sample.

Assumption 2. I shall next assume that *corresponding tests in the two batteries may be treated as* ' *equivalent* '. Burt's definition of ' equivalence ' differs somewhat from that ordinarily put forward, and is perhaps a little too stringent. For any pair of tests to be regarded as ' equivalent ' (he says), (i) their means and (ii) their standard deviations must be the same ; (iii) they must depend on the same common factors (whether general and bipolar or basic and group) ; (iv) they must have the same weight for the same common factors and the same weight for the error factors (the error factors, of course, unlike the common factors, will not be the same).[1] From this it would follow that the corresponding self-correlations and intercorrelations in the N.W. and S.E. quadrants respectively differ only by such small amounts as may be attributable to errors of sampling. Evidently this is a corollary whose validity can be tested for any given set of data. In the present research, it will be seen, some of the observed intercorrelations do differ significantly ; and these differences indicate obvious points at which the batteries require, if possible, to be improved. However, the totals for the two quadrants (12·222 and 12·308) differ hardly at all ; and that may perhaps suffice to justify a provisional acceptance.

With this assumption the best estimate for the common standard deviation of the two batteries will be obtained by taking the arithmetic mean of the variances of both ; and accordingly, as Burt suggests, we may now use the formula for the *intra-class correlation* (instead of the more familiar product-moment formula) together with the scheme for analysing variance associated with this coefficient.[2] In place of eqn. (ii *a*), therefore, we shall for the future substitute

$$r_{ss'} = \frac{\sum R_{ab}}{\frac{1}{2}(\sum R_{aa} + \sum R_{bb})} \qquad \text{(viii)}$$

With the present data, the two sums, $\sum R_{aa}$ and $\sum R_{bb}$, are so nearly equal that the substitution of the arithmetic mean for the geometric leaves the value for $r_{ss'}$ (·8205 as given above) virtually unchanged.

Assumption 3. The summations carried out in an analysis of variance give equal weights to each component of the same type : in effect, the method assumes that $g_{11} = g_{21} = \ldots = \bar{g}$ (say), and $d_{11} = d_{21} = \ldots = \bar{d}$ (say), and so on, where $\bar{g}$, $\bar{d}$, etc., denote the average saturations. It is, of course, a familiar principle that, in summing a number of test-measurements, unless the differences between the weights are large, the substitution of equal weights has little effect on the relative values of the sums obtained (see this *Journal*, III, p. 111). Moreover, in constructing two parallel batteries of tests designed to measure the same ability, the investigator will, as a rule, discard any subtests which have a decidedly low weight for the general factor, or are likely to be more influenced than the rest by the conditions peculiar to a given trial. We may, therefore, now introduce this third assumption, which will still further simplify the resulting formulae and bring our factorization still more into line with the analysis of variance, namely, that, though the weights (or saturations) for different factors will usually be different, *the weights for the same factor are the same for all the tests.*

In addition then to what may be called the simple ' four factor hypothesis ', our model will for the present assume (i) that the factors postulated are orthogonal, (ii) that as regards their factorial composition the test-batteries are equivalent, and (iii) that the saturations may differ from one factor to another but are the same for the same factor. In our present data there is nothing that flagrantly conflicts with these assumptions : but for a strict test of their accuracy a full factor analysis would be essential. This we shall attempt at a later stage.

We can now reconstruct the correlations to be expected on the basis of these last two assumptions. As will be seen from Table IV, the hypothesis on which our model is based treats the N.W. and S.E.

TABLE IV. POOLING SQUARE WITH EQUALIZED SATURATIONS

Sums of Correlations in Terms of Hypothetical Saturations

Factor Saturations	Summed Correlations	
$\bar{g}\ \bar{d} - \bar{t} - \bar{e} - - -$ $\bar{g}\ \bar{d} - - \bar{t} - \bar{e} - -$ $\ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots$	$\sum R_{aa} = (n\bar{g})^2 + (n\bar{d})^2 + n\bar{t}^2 + n\bar{e}^2$	$\sum R_{ab} = (n\bar{g})^2 + n\bar{t}^2$
$\bar{g} - \bar{d}\ \bar{t} - - - \bar{e} —$ $\bar{g} - \bar{d} - \bar{t} - - - \bar{e}$ $\ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots\ \ldots$	$\sum R_{ba} = (n\bar{g})^2 + n\bar{t}^2$	$\sum R_{bb} = (n\bar{g})^2 + (n\bar{d})^2 + n\bar{t}^2 + n\bar{e}^2$

[1] Once again I have ventured to simplify. Burt distinguishes between ' equivalent tests ' and ' parallel tests '. But this is unnecessary for our present purpose.

[2] See Fisher, *loc. cit. sup.*, pp. 199 f., and Tables 38 and 39.

quadrants as identical not only in their sums but also in their composition. Accordingly, since we now have $R_{bb} = R_{aa}$, we can write

$$r_{ss'} = \frac{\sum R_{ab}}{\sum R_{aa}} = \frac{(n\bar{g})^2+n\bar{t}^2}{(n\bar{g})^2+(n\bar{d})^2+n\bar{t}^2+n\bar{e}^2} = \frac{A}{A+B}, \quad \therefore \quad \text{(ix)}$$

in Fisher's notation. Thus interpreted, the ordinary reliability measures *the ratio of* (*a*) *the variance resulting from the general factor and the factors specific to similar tests to* (*b*) *the total variance, which includes in addition the variance due to the different occasions and to the errors of measurement.*

Is this what we really want to measure when we are estimating the reliability of our battery?

Computation of Variances. Before we take up this question, it will be well to consider whether it is really possible to estimate the values of these averaged factor saturations, and if so how. As a matter of fact two alternative procedures are available. With standardized measurements the *average inter-correlation* for the separate tests is identical with the *intra-class correlation*, and an unbiased estimate of this can readily be derived from the variances.[1] Hence we can start either from the variances or from the detailed correlations.

(a) *The Factorial Analysis of Variance.* From the symbolic tables set out above, it is obvious that what is wanted are not the factor saturations as such, but merely their squares, i.e., the factor variances. These can readily be computed by the device which Burt has called the 'factorial analysis of variance'—a further analysis in which the crude variances, computed for an ordinary analysis of variance, are themselves re-analysed to obtain, as it were, the purified variances for the several hypothetical factors that express the classification employed.

The requisite formulae are given in the earlier memorandum already cited. Using V to denote a crude variance derived by an ordinary analysis of variance (as in Table 1 above) and σ^2 to denote the variances of the hypothetical factors, we have the following equations, and then, inserting the values for V given in Table 1, we obtain the figures shown on the right:

$\sigma_p^2 = (V_p - V_{pt} - V_{po} + V_{pto}) \div mn$	·5811
$\sigma_{pt}^2 = (V_{pt} - V_{pto}) \div m$	·1916
$\sigma_{po}^2 = (V_{po} - V_{pto}) \div n$	·1077
and $\sigma_{pto}^2 = V_{pto}$	·1196
$\sigma_p^2+\sigma_{pt}^2+\sigma_{po}^2+\sigma_{pto}^2 = V_s$	1·0000

It will be remembered that the marks for each subtest were reduced to standard measure. The square-sum for a standardized subtest is $N-1$ (the number of degrees of freedom in the sample of persons), i.e., with the present sample 86. Since there are $m \times n = 8$ subtests, the total square-sum must be $86 \times 8 = 688$. But this was also the figure obtained for Q_t by adding the sums of squares in column 1 of Table I. Re-dividing that total by the number of degrees of freedom $(N-1)mn$, we obtain for the average 'mean square' of any one measurement the value of 1·000 (the figure for V_s in the table above). These results provide a useful check on the computer's arithmetic.

The four values given in the table above show how this average variance may itself be subdivided to show the proportionate contributions from the four hypothetical factors. The basic factor contributes over 58 per cent.—a decidedly encouraging result; the supplementary factor for tests contributes over 19 per cent.—an unduly large amount; and the two smaller factors, representing susceptibility to the influences of the occasion and error of measurement in the narrower sense, together contribute another 22 per cent. or rather more. Evidently, before attempting to revise the test material, it will be desirable to investigate more closely the precise causes of these undesirable fluctuations: to that we shall turn in a moment.

(b) *Correlational Factors.* From the formulae already summarized in Table IV we can deduce the relations between the factors obtained by this 'further analysis of variance' and those obtained by the 'sum method' of factorizing the detailed correlations.[2] The method of deduction will be

[1] Cf. C. Burt, *Factors of the Mind*, 1940, p. 275, and the I.I.E.C. memoranda referred to above.
[2] Cf. this *Journal*, II (1949), 62 (appendix).

simplified if, as Table IV suggests, we first average the intercorrelations for the several subte sts We can thus re-write the values in the pooling square as follows :

			Total				Total
1	$\bar{r}_s$	...	$1+(n-1)\bar{r}_s$	$\bar{r}_{sd}$	$\bar{r}_d$	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$
$\bar{r}_s$	1	...	$1+(n-1)\bar{r}_s$	$\bar{r}_d$	r_{sd}	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$
...	...	...	...	...	...	...	...
Total			$n[1+(n-1)\bar{r}_s]$				$n[\bar{r}_{sd}+(n-1)\bar{r}_d]$
$\bar{r}_{sd}$	$\bar{r}_d$	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$	1	$\bar{r}_s$	...	$1+(n-1)\bar{r}_s$
$\bar{r}_d$	$\bar{r}_{sd}$	...	$\bar{r}_{sd}+(n-1)\bar{r}_d$	$\bar{r}_s$	1	...	$1+(n-1)\bar{r}_s$
...	...	...	...	...	...	...	etc.

Or, substituting the numerical averages from Table I,

			Total				Total
1·0000	·6888	...	3·0664	·7727	·5811	...	2·5160
·6888	1·0000	...	3·0664	·5811	·7727	...	2·5160
...	...	...	...	...	...	...	...
—	—	—	12·2656	—	—	—	10·0640
...	...	...	...	...	...	...	...

If (as here) the tests have been standardised so that $\sum\sigma^2 = 1{\cdot}000$, then the values for the factor variances will be given by the following equations :

$$\bar{g}^2 = \sigma_p^2 = \bar{r}_d = {\cdot}5811,$$
$$\bar{t}^2 = \sigma_{pt}^2 = \bar{r}_{sd}-\bar{r}_d = {\cdot}7727-{\cdot}5811 = {\cdot}1916,$$
$$\bar{d}^2 = \sigma_{po}^2 = \bar{r}_s-\bar{r}_d = {\cdot}6888-{\cdot}5811 = {\cdot}1077,$$
$$\bar{e}^2 = \sigma_{pto}^2 = 1-\bar{r}_{sd}-\bar{r}_s+\bar{r}_d = 1-{\cdot}7727-{\cdot}6888+{\cdot}5811 = {\cdot}1196.$$

It will be seen that these equations yield exactly the same values as before. They are in fact merely a simplification of the ordinary method of group factor analysis, adapted to fit the hypothetical scheme of basic, group, and specific factors that we postulated at the outset.

Various Forms of the Reliability Coefficient. Now, as has frequently been pointed out, there are in practice several alternative ways of estimating the socalled reliability of a test ; and these different procedures are by no means equivalent. Indeed, they imply quite different definitions for what is meant by reliability ; and much confusion would be avoided if they were given different names. Three main forms [1] are commonly distinguished ; and, as Burt has observed, they may most conveniently be expressed in terms of factors such as we have just described. Slightly modifying his definitions to bring them into line with those of later writers, we may formulate them as follows.[2]

[1] E.g., J. P. Guilford, *Psychometric Methods* (1936), p. 411 ; H. Gulliksen, *The Theory of Mental Tests* (1950), pp. 193 f.

[2] One of the earliest publications which uses factor-theory in a discussion of reliability is the article by Miss B. M. D. Cast on marking English composition (*Brit. J. Educ. Psychol.*, IX, 1939, pp. 257 f., X, 1940, pp. 49 f.). Her procedure has been largely followed in Ebblewhite-Smith, *Marking English Essays* (1941), pp. 42 f. ; and was itself based on Burt's earlier L.C.C. Memorandum. Pilliner has similarly adopted Burt's ' three-way analysis of variance ' to describe the correlations between marks awarded by different markers to different essays written by the same pupils ; but apparently he does not accept the use of the intra-class correlation, and he retains the differences between the means for essays and the means for marks (A. E. G. Pilliner, ' Applications of Analysis of Variance to Problems of Correlation ', this *Journal*, V, 1952), pp. 31-8). H. W. Alexander, like Burt, makes full use of the intra-class correlation (' The Estimation of Reliability when Several Trials are available ', *Psychometrika*, XII, 1947, pp. 79-100). His variance equations for the biased and unbiased forms are the same as Burt's (cf. *Factors of the Mind* (1940), pp. 274 f.) and so is his formula showing the relation between the intraclass correlation and the average correlation.

An instructive discussion of the problems raised, which is also expressed in terms of factor analysis (viz., Holzinger's bifactor theory) will be found in L. J. Cronbach, ' Test Reliability : Its Meaning and Determination ' (*Psychometrika*, XII, 1947, pp. 1-16). The algebraic expressions there proposed are not explicitly related to the techniques of the pooling square or the analysis of variance : and the factors refer

Definition 1. What Burt calls the ' Coefficient of Precision ' is that form of reliability coefficient which measures the degree to which the measurements obtained from a test or battery on one occasion will agree with the measurements obtained from the same test or battery on a second occasion. This was the original meaning intended by those early psychologists who introduced the term. Such a conception tacitly assumes (*a*) that the function tested itself remains unaltered during the interval of time that separates the two testings; (*b*) that on the second occasion the measurements or the functions are themselves unaffected by the fact that the function has already been tested on a previous occasion; and (*c*) that each subtest measures the same function as the battery taken as a whole, i.e., that ' unique ' or ' specific factors ' (other than those that express accidental influences) have been eliminated either experimentally or statistically. These assumptions imply that both $\bar{t}^2$ and $\bar{d}^2$ will be zero, or at any rate negligible, or, what amounts to the same thing, that both the interaction between persons and tests and the interaction between persons and occasions are non-significant. In that case the expression for $r_{ss'}$ would reduce to

$$r_p \text{ (say)} = \frac{n\bar{g}^2}{n\bar{g}^2+\bar{e}^2} = \frac{n\sigma_p^2}{n\sigma_p^2+\sigma_{pto}^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(x)}$$

This simple ' two factor ' interpretation was implicitly accepted by Spearman when introducing his formula for ' correcting ' validity-coefficients for the unreliability of the two sets of assessments. The implied analysis is clearly brought out in the proof of Spearman's formula given by Yule, who at the same time indicated methods for testing the legitimacy of the main assumptions.[1] Much the same assumptions are regularly made by the physicist when he assesses the precision of his measurements by repeating his observations; but with psychological assessments the three assumptions of stability, of independence, and of non-specificity may each or all be gravely violated by the data actually obtained. This demands obvious modifications in the psychologist's proceedure.

(i) To begin with, since the two batteries are applied on two distinct occasions, it is always possible that the ability may have changed, or that its manifestations on the second occasion may differ from its manifestations on the first, either because the conditions on each occasion are different, or because the previous testing may have produced practice-effects, or else because, to avoid memory and practice, different versions of the tests have been applied. Whenever there are any prior reasons for anticipating some such influence, we must re-introduce the component represented by $\bar{d}^2$. We then obtain

$$r_{p'} = \frac{n\bar{g}^2}{n\bar{g}^2+n\bar{d}^2+\bar{e}^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xi)}$$

The effect of the additional component will obviously be to *reduce* the apparent reliability; Spearman's correction formula would therefore exaggerate the ' true ' value of the validity coefficient.

(ii) When the subtests used on the second occasion are the same as, or at least similar to, those used on the first, it will be natural to expect an increase in the correlations between similar subtests, either as a result of specific factors or possibly of practice. This will require the re-introduction of the component whose variance is represented by $\bar{t}^2$, a component common to both occasions. If at the same time we assume (as Spearman does) that the conditions on the two occasions are identical, the formula will be

$$r_{p''} = \frac{n\bar{g}^2+\bar{t}^2}{n\bar{g}^2+\bar{t}^2+\bar{e}^2} \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xii)}$$

This time the effect of the additional component will be to *increase* the apparent reliability.

There is no satisfactory way of deriving $r_{p'}$ or $r_{p''}$ from the pooling square. When the interactions omitted from the formulae are really non-significant, tentative estimates could be deduced from a ' factorial analysis of variance '. But these simplified versions of the reliability coefficient will seldom be wanted in a systematic research where adequate data have been secured. Consequently, they call for no special name or formal definition. In practice, it will nearly always be safer to assume that *both* these irrelevant disturbances may be operative, even if their effects do not seem fully significant. Consequently, the expression for the total variance, which forms the denominator of the ratio,

---

to ' items ' within a single test, not to subtests comprising a battery. Consequently the treatment is somewhat more complicated, e.g., the basic equation contains seven terms, five of which are themselves sums. On the other hand, Cronbach's equations do not explicitly contain factors for the different occasions or days. Moreover, whereas Burt distinguishes types of reliability coefficient in terms of the factors *common* to two (or more) applications, Cronbach distinguishes his types by specifying the factors regarded as sources of *error*.

[1] C. Spearman, ' Proof and Measurement of Association between Two Things ', *Amer. J. Psychol.*, XV, 1904, p. 88; G. U. Yule, *Introduction to the Theory of Statistics* (1912), pp. 213-14, sect. 7.

must allow for the presence of both additional sources of variation. But are we also to include them in the numerator of the ratio, or should we endeavour to eliminate them? The answer depends on what precisely we desire to measure. At least three different replies are possible. Hence we have to consider three further types of reliability coefficient and three further formulae.

Definition 2. The ' Coefficient of External Consistency ' or ' Equivalence ' measures the degree to which a composite test, applied on a second occasion, either in the same or in a closely parallel form, yields results agreeing with those of the first application, even though the conditions of testing may differ. If the tests are identical, we may speak of ' consistency ' ; if not, of ' equivalence '.[1] For this purpose we evidently require to introduce $\bar{t}^2$ into the numerator and both $\bar{t}^2$ and $(n\bar{d})^2$ into the denominator. The formula will therefore be that already reached (eqn. ix) when calculating a test-retest correlation from the pooling square, namely:

$$r_e = \frac{n\bar{g}^2+\bar{t}^2}{n\bar{g}^2+n\bar{d}^2+\bar{t}^2+\bar{e}^2} = \frac{n\,\sigma_p^2+\sigma_{pt}^2}{n\,\sigma_p^2+n\,\sigma_{po}^2+\sigma_{pt}^2+\sigma_e^2}. \quad . \quad . \quad . \quad \text{(xiii)}$$

This is the form to which eqn. ix reduces when n is cancelled from both numerator and denominator. We may therefore regard it as the variance equivalent of the familiar ' test-retest correlation ' with parallel forms.

Substituting the values already obtained, we have

$$r_e = \frac{4\times\cdot5811+\cdot1916}{4\times\cdot5811+4\times\cdot1077+\cdot1916+\cdot1196} = \frac{2\cdot5160}{3\cdot0664} = \cdot8205.$$

Or, using the average correlations [2] as they stand,

$$r_e = \frac{\bar{r}_{sd}+(n-1)\bar{r}_d}{1+(n-1)\bar{r}_s} = \frac{2\cdot5160}{3\cdot0664} = \cdot8205.$$

Nevertheless, instructive as this coefficient may be for preliminary inquiries, it does not appear to furnish precisely the information we require. Our tests have been designed to measure the common ability, g (whose variance is designated by σ_p^2) : the more specific abilities peculiar to the several subtests (whose variance is designated by σ_{pt}^2) do not, as a rule, form part of the particular function we set out to test.[3] We must therefore seek some way of eliminating these disturbing factors from the numerator. If we keep to a correlational technique, the appropriate device seems plain. Much the same difficulty confronted the earlier factorists ; and to eliminate the irrelevant specifics from their correlation matrices, they proceeded to substitute ' reduced self-correlations ' or reduced self-covariances ' for the self-correlations or observed test-variances entered in the leading diagonal. Here we can apply the same procedure either (i) to the submatrix of cross-correlations (R_{ab}), thus eliminating $\bar{t}^2$, or (ii) to the submatrix of intercorrelations (R_{aa}), thus eliminating $(\bar{t}^2+\bar{e}^2)$. In ordinary factor analysis the values to be inserted would be reached by successive approximation ; but if (as we are here assuming) the saturations for the same factor do not differ greatly in magnitude, we may plausibly treat the average value of the self-correlations as approximately equal to the average of the inter-correlations. To indicate that the submatrices now have reduced values in their diagonals, I shall follow the usual convention, and affix an asterisk.

By referring to Tables I and IV, the reader can readily verify for himself the fact that the correlational procedure, with these reductions, yields the same values as the variance procedure, with the proposed omissions. We have in fact

$$\begin{aligned}
\Sigma R_{aa} &= n(n\sigma_p^2+n\sigma_{po}^2+\sigma_{pt}^2+\sigma_e^2) = 4(2\cdot3244+\cdot4308+\cdot1916+\cdot1196) = 12\cdot2656\,;\\
\Sigma R^*_{aa} &= n(n\sigma_p^2+n\sigma_{po}^2) = 4(2\cdot3244+\cdot4308) = 11\cdot0208\,;\\
\Sigma R_{ab} &= n(n\sigma_p^2+\sigma_{pt}^2) = 4(2\cdot3244+\cdot1916) = 10\cdot064\,;\\
\Sigma R^*_{ab} &= n(n\sigma_p^2) = 4(2\cdot3244) = 9\cdot2976.
\end{aligned}$$

[1] The term ' coefficient of consistency ' was proposed by P. Hartog (*Marks of Examiners* (1936), p. xiv, footnote 1). The phrase ' coefficient of *internal* consistency ' was suggested to denote more specifically the coefficient derived from data obtained at a single administration, i.e., from the intercorrelations of the subtests, and the phrase ' coefficient of *external* consistency ' to denote the coefficient derived from two administrations.

[2] Cf. Burt, *Marks of Examiners*, p. 303, eqn. xl.

[3] In some inquiries, particularly in the field of educational and vocational selection, the examiner may desire to test *both* general capacity *and* special aptitudes or knowledge. Thus, in the old junior county scholarship examination, tests involving reading, spelling, composition, and arithmetic were set with the intention of assessing not only the candidates' general educational ability (g) but also their special scholastic abilities and attainments (t). In such cases eqn. xiii remains the appropriate formula.

As a result of these various modifications, we obtain three further coefficients.

Definition 3. The ' Coefficient of Internal Consistency ' measures the degree to which the results of the various *constituent* tests agree with each other, and therefore the degree to which the results of a second application, with a battery assumed to be internally identical with the first and under conditions assumed to be exactly the same, would agree with the results of the first. On referring to the symbolic correlation matrix (Table IV) it is evident that the theoretical formula required by the above assumptions will be

$$r_c = \frac{n(\bar{g}^2+\bar{d}^2)}{n(\bar{g}^2+\bar{d}^2)+\bar{t}^2+\bar{e}^2} = \frac{\sum R^*_{aa}}{\sum R_{aa}} = \frac{11{\cdot}0208}{12{\cdot}2656} = {\cdot}8985. \quad . \quad . \quad . \quad . \quad \text{(xiv)}$$

The coefficient refers primarily to the internal consistency of the test as assessed by the results of a single trial.[1] But naturally, the figure will vary somewhat with the results of different trials. Here we have in effect averaged the consistency coefficient for two. If there is any likelihood that practice or familiarity may alter the internal consistency, then we must base our coefficient solely on the first trial. For that, the same procedure here yields $r_c = {\cdot}8970$.

The equation can thus be employed when the tests have only been administered once ; this of course is the attraction of all such formulae. However, with a single trial it is obviously impossible to distinguish between $\bar{g}^2$ and $\bar{d}^2$ or between $\bar{t}^2$ and $\bar{e}^2$: the two former are merged into a single factor common to all the four tests and the two latter represent the specific factors peculiar to the results obtained from a single test on a single occasion. Consequently the pooled variances may be re-written

$$\frac{\sigma^2_G}{\sigma^2_G+\sigma^2_E}; \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xv)}$$

and this expression, as Burt has shown, is equivalent to the ' split half ' reliability coefficient obtained from a single trial *when all possible splits are taken and averaged* (this *Journal*, VI, 61).

Definition 4. The ' Coefficient of Test Stability ' indicates the degree to which the measurements of an ability assessed by one set of tests remain the same (or would remain the same) when measured by a second set, even though the methods or conditions of testing may differ. The theoretical formula will be

$$r_{s/t} = \frac{n\bar{g}^2}{n(\bar{g}^2+\bar{d}^2)+\bar{t}^2+\bar{e}^2} = \frac{\sum R^*_{ab}}{\sum R_{aa}} = \frac{9{\cdot}2976}{12{\cdot}2656} = {\cdot}7580. \quad . \quad . \quad . \quad \text{(xvi)}$$

Assuming that g (the general factor as assessed by the unweighted means for the several persons) represents the characteristic that the battery of tests was designed to measure, this equation comes nearest to expressing what most psychologists appear to have meant by reliability. It will be seen that the formula can only be computed if data from at least two trials are available. The tests should be so compiled, and the administration so carried out, that both $\bar{d}^2$ and $\bar{t}^2$ are reduced to a minimum ; and, when that has been done, the coefficient so calculated will probably approximate to the coefficient of internal consistency. But, unless the structure of the test and the peculiarities of the function tested (e.g., its liability to change) are already well known, it would be rash to state whether that coefficient indicates a lower or an upper bound.

Definition 5. The ' Coefficient of Person Stability ' measures the extent to which the relative abilities of a given set of persons, assessed on two or more separate days, have remained the same, in spite of the interval between the two applications or (particularly if the interval is short) in spite of the variations in the conditions that obtained. For this the most obvious formula will be

$$r_{s/p} = \frac{\bar{g}^2}{\bar{g}^2+\bar{d}^2} = \frac{\sum R^*_{ab}}{\sum R^*_{aa}} = \frac{9{\cdot}2976}{11{\cdot}0208} = {\cdot}8437. \quad . \quad . \quad . \quad . \quad \text{(xvii)}$$

This concept is not unlike that designated ' function stability ' by certain writers.[2] But the explanations generally offered overlook an important qualification. A given function or ability may tend to change in *all* persons at approximately the same absolute or proportional rate : if, for example,

[1] " The determination of reliability from a single test-application " is fully discussed by Burt in ' The Reliability of Teachers' Assessments ', *Brit. J. Educ. Psychol.*, XV, 1945, pp. 80-92, and the use of analysis of variance illustrated in detail.

[2] Cf. G. B. Paulsen, ' A Coefficient of Trait Variability ', *Psychol. Bull.*, XXVIII, 1931, pp. 218 f., and R. H. Thouless, ' Test Unreliability and Function Fluctuation ', *Brit. J. Psychol.*, XXVI, 1936, pp. 325 f. I should add that the more important formulae given above are virtually identical with those suggested by Burt in his memorandum for the International Institute Examinations Inquiry : cf. *Marks of Examiners* (1936), p. 274, eqns. xii and xv. They are here given with a slightly different subscript-notation adopted in his *Laboratory Notes*.

the I.Q. were absolutely constant for every child, the mental ages of the children tested would alter, and alter by different annual increments, but the spaced order of merit would remain the same from year to year; the same would happen if all pupils increased in the knowledge of a given school subject in accordance with some constant educational quotient. This kind of change is not included in the foregoing concept. In actual fact, however, as numerous investigations have shown, the correlations between the I.Q.'s of the same group of pupils tend to diminish with increase in time; and this implies that the changes in mental age take place more or less erratically, so that the order of ability does not remain the same. Hence what we require to measure is the *relative stability of the persons* rather than the stability of the function or capacity in and for itself.

The Coefficient of Trait Variability. By subtracting any of the foregoing ' coefficients of reliability ' from unity, we could if we wished obtain a ' coefficient of unreliability ' (or ' error '). There is, however, one additional coefficient of unreliability that deserves special mention, namely, the ' Coefficient of Trait Variability '. This measures the degree to which the trait or ability fluctuates from one occasion to another. The amount of variation is indicated by subtracting the ' Coefficient of Test Stability ' (i.e., stability of the test measurements from one day to another), not from unity, but from the ' Coefficient of Internal Consistency ' (i.e., consistency of tests applied on the same day). This yields

$$r_{(tv)} = \frac{n\sigma^2_{po}}{n\left(\sigma^2_p+\sigma^2_{po}\right)+\sigma^2_{pt}+\sigma^2_{pto}} = \cdot 8985-\cdot 7580 = \cdot 1405. \quad . \quad . \quad . \quad \text{(xviii)}$$

It is instructive to compare this with Thouless's formula for ' function fluctuation ' (*loc. cit.*, p. 332),

$$I_{(ff)} = \frac{r_{(a_1-a_2)(b_1-b_2)}}{\frac{1}{2}\left(r_{a_1b_1}+r_{a_2b_2}\right)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xix)}$$

where a and b denote the two tests, and 1 and 2 the occasions on which they are applied. It is not easy to express this in terms of the constants we have so far used, because with only two tests and two occasions the variances of the tests and still more of their differences are liable to be highly erratic. However, making the same assumptions as before,[1] it becomes very approximately equivalent to

$$\frac{\sigma^2_{po}}{\left(\sigma^2_{po}+\sigma^2_{pto}\right)\left(\sigma^2_p+\sigma^2_{po}\right)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(xx)}$$

[1] Burt points out that Yule (in his discussion of attenuation due to unreliability) first suggested calculating the correlation $r_{(a_1-2)(b_1-b_2)}$ as a ' partial test ' for the presence or absence of ' fluctuation ', as later writers have called it (U. Yule, *Introduction to the Theory of Statistics* (1910), p. 214; Yule's method was used by Brown to compute the ' variability of function within the individual ' from his own data, *Essentials of Mental Measurement* (1911), p. 84). But, as Burt observes, ' the precise grounds for the choice of the denominator (for eqn. xix) are by no means clear '. With the usual assumptions regarding equality of variances and independence of factors in the population (assumptions which may be quite flagrantly vitiated in small samples and with factors derived from unweighted means) he goes on to show that Thouless's coefficient is approximately equivalent to

$$\frac{(V_{po}-V_{pto})}{(V_{po}+V_{pto})} \cdot \frac{(V_p+V_{pt}+V_{po}+V_{pto})}{(V_p-V_{pt}+V_{po}-V_{pto})} \quad . \quad . \quad . \quad . \quad . \quad \text{(xx } a\text{)}$$

which reduces to equation xx when (as here) the measurements are so standardized that

$$\sigma^2_p+\sigma^2_{pt}+\sigma^2_{po}+\sigma^2_{pto} = 1{\cdot}00.$$

In his worked example he uses the constants obtained in his article on ' Factor Analysis and Analysis of Variance ' (this *Journal*, I, p. 12) and obtains (a) with Thouless's formula ·031/·608 = ·051 and (b) with eqn. xx a ·038/·617 = ·062. It will be observed that eqn. xx, or its equivalent eqn. xx a, yields a value of zero when σ^2_{po} is zero and a value of unity when Vp and Vpt are both zero. This version of the ratio would therefore seem better than the correlational form, but even so is not quite consistent. However, ' for any genuine assessment of the amount of instability or trend in the ability or " function " tested, *more* than two trials are essential; and the foregoing formulae are accordingly generalized to meet cases that involve more than two trials and more than two subtests in each trial. Where the trends are linear (whether produced by positive or negative practice effects due to the tests themselves or by mental growth or training), the ordinary methods of analysis of variance can be employed; if they are curvilinear or irregular, the line for trend can be fitted by the methods discussed by Philpott and Burt in their investigations of work-curves. But, for a really adequate investigation of the question, factorial methods, such as are used in the later part of the article, seem indispensable.

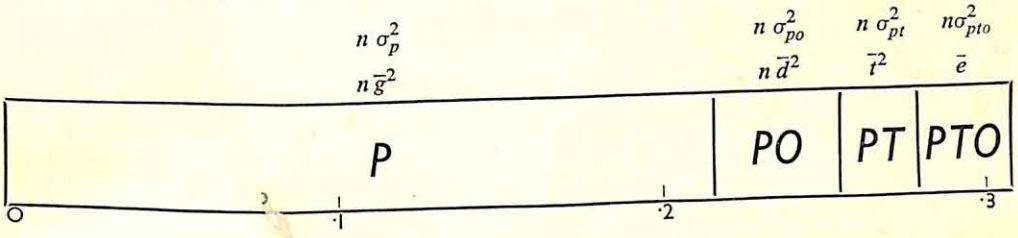
While the present paper is passing through the press, an article by C. C. Anderson on ' Simple Methods of Testing for Function Fluctuation ' has appeared in *Brit. J. Psychol.*, XLVI, 1955, pp. 1-19. His ' model ' is similar to Burt's, and from it he derives a formula for ' Thouless's criterion ' in terms of ' mean squares '.

With the foregoing figures eqn. xx yields a value of 0·68 ; eqn. xix (after similar subtests have been pooled) yields one of 0·61. Both values seem far too high.

Comparative Values of the Various Coefficients. It is often supposed that the different methods of assessing reliability may be arranged in order according to the size of the coefficients they yield. For instance, several writers have stated that estimates based on single trials—estimates of ' internal consistency '—are higher than those based on two trials—the ' test-retest ' procedure : the former it is said, avoids the attenuation due to different days and different conditions, but may be unduly enlarged when irrelevant influences operative on the same day affect *all* the subtests ; the latter is diminished by the varying changes the examinees may undergo in the interval between the trials ; and, if ' equivalent forms ' have been employed, will be still further reduced by the lack of complete equivalence. Other writers have drawn the opposite conclusion. Jackson, for example, compiled five tests of intelligence, and found in every case that estimates based on test and retest were higher than those based on internal consistency (*Psychometrika*, VII, 159) : such a result is to be expected when the compiler selects a heterogeneous set of problems in order to assess intelligence from different angles and perhaps deliberately discards those that are likely to duplicate each other.

A comparison of the foregoing formulae will show that inconsistencies like these are almost inevitable. The differences between one coefficient and another do not result from just adding yet another element in turn to the ' true variance ' : sometimes one element is substituted for another in the

FIG. 1.—Contributions of Component Variances to Total Variance.



numerator, and others are omitted or inserted in the denominator. In Figure 1 the magnitude of the four component variances as obtained in the present experiment is illustrated by a ' bar diagram '.[1]

Unfortunately he gives no proof, and there are at least two misprints in the formula. Instead of what is there printed, we should probably read (in Anderson's notation) $\frac{B-C}{B+(p-1)C}$. But since in all his experiments $p = 2$, this simplifies to $\frac{B-C}{B+C}$. Anderson's formula is thus equivalent to omitting the division by ·617 in eqn. (*b*) above, and seems to assume that the denominator in Thouless's ratio may be taken as approaching unity when the factors are genuinely independent, and as counteracting the effect of their correlations when they are not.

An alternative formula suggested by Burt is

$$I_{(tv)} = \frac{\sum R_{aa} - \sum R_{ab}}{n^2 + \sum R_{ab}} = \frac{n\sigma_{po}^2 + \sigma_{pto}^2}{n(\sigma_{po}^2 + \sigma_{pto}^2) + (n-1)\sigma_{pt}^2}$$

If the variance due to *occasions* is zero, this reduces to $\sigma_{pto}^2/[n\sigma_{pto}^2 + (n-1)\sigma_{pt}^2]$, which diminishes indefinitely as n increases indefinitely, and in any case becomes exactly zero if the error variance is also zero. If the specific variance due to the *tests* is zero, it reduces to $(n\sigma_{po}^2 + \sigma_{pto}^2)/(n\sigma_{po}^2 + n\sigma_{pto}^2)$, and becomes exactly unity if the error variance is also zero. With the present data the value of this ' index ' would be ·3710—a far more plausible figure.

The chief differences between $r_{(tv)}$ and $I_{(tv)}$ arise from the addition of the error variance in the numerator and the disappearance of the general factor variance from the denominator. This seems plausible, since the absolute size of the general factor variance is quite irrelevant to the question raised which the ' index ' seeks to answer.

[1] The diagram is intended to represent a simple and instructive piece of apparatus used for demonstrations by Burt. A long and narrow tray (similar in form to the diagram) holds a sequence of four oblong panels, painted white on the front and black on the back. When the constituents are included in the numerator (the ' true variance '), the white side is left uppermost ; when they are included only in the denominator (the ' total variance ') and thus represent part of the ' error variance ', they are turned over and the black side is uppermost ; when they are excluded from both, they are removed altogether. Various sets of four such panels represent the results obtained from different types of experiment.

From this it will at once be seen that with the present data the substitution of σ^2_{pt} for $n\sigma^2_{po}$ will reduce the value; with other data it may have the opposite effect; and similarly with the changes in the denominator. In short, the equations we have reached make it obvious that no fixed order of magnitude, valid for all kinds of data, can possibly be laid down.

Factor Analysis. The formulae which I have so far described may suffice the purposes of a preliminary inquiry. It must, however, be remembered that they were based on certain simplifying assumptions, introduced partly to permit the adoption of well recognized routine techniques, such as an ordinary analysis of variance, and partly to keep the working methods as short and easy as possible. Let us now examine the effect of removing some of the most questionable assumptions, and inquire how far the more elaborate modes of computation will alter the main results.

Throughout the foregoing calculations we have assumed that the weights (i.e., the 'loadings' or 'saturations') for any one factor may be treated as equal for all the tests. This assumption, as we have seen, is implicit in the formulae for the analysis of variance. Yet in practice it is seldom likely to be strictly true; and it is certainly conceivable that the appropriate weights might differ so widely as to render the inferences hitherto drawn decidedly precarious. The matrix representing the interaction between tests and persons itself, as was noted at the time, also deserves a more detailed treatment. Our simplified procedure assumed that the factors producing this interaction were merely isolated components specific to each test—in other words, that, apart from the figures in the diagonal, the correlational submatrices have unit rank.[1] But once again such an assumption may be quite unwarranted. If so, it becomes conceivable that our previous estimate for the reliability of the battery as a test of intelligence will turn out to have been unduly magnified by the inclusion of more specialized abilities which we had no intention of measuring. The obvious way to meet both these difficulties is to subject the data to a formal factor analysis.

In our own work we have followed the procedures recommended and used by Dr. Barakat and Dr. Moursy in their recent papers.[2]

(a) *With Bipolar Factors.* The matrix of observed correlations, given above in Table II, was first factorized by 'simple summation'. Three successive approximations (with incidental adjustments) sufficed to secure appropriate values for the reduced self-correlations. With an 8×8 matrix, the observed coefficients will provide 28 degrees of freedom; four summational factors would account for all but two. Four were accordingly extracted; but the last was too small to be statistically significant, and seemed devoid of any assignable meaning. The saturations for the three largest are shown in Table V.

(I) The first is a general factor, contributing over 60 per cent. to the total variance. It would presumably be identified, by those who prefer a summational procedure, with the general ability which the tests were designed to measure. The correlation between the factor-measurements for the individual testees and the independent assessments for intelligence provided by the teachers amounted to 0·67. This is certainly an improvement on the value obtained by taking an unweighted sum of the marks for the several tests (0·59).

(II) The second factor contributes more than 10 per cent. to the total variance; and was quite unexpected in its nature. It classifies each battery into two distinct parts. We at first supposed this might result from differences in the material used for the four subtests. For the first two it consisted of coloured blocks or insets to be manipulated by the child; for the second two, of black-and-white patterns on paper. But, after securing introspections from some of the children, and trying similar tests with students, we found that the material in which the problems were embodied had far less influence than the type of activity they involved: in the first two subtests the dominant process was apparently perceptual combination or construction; in the other two it was perceptual discrimination: in short, the first two seemed to depend largely on mental synthesis, the latter on mental analysis.[3]

(III) The third factor reveals an obvious contrast between the results of the first and second trials respectively. With this method of analysis it contributes rather less to the total variance than might have been anticipated, namely, 5·5 per cent.

In the light of Burt's comparison between the results of an ordinary analysis of variance and those of a summational analysis (this *Journal*, I, 12 and 25, Tables VI and XV), we were originally tempted to identify the three foregoing factors with the three components which nearly always underlie the analysis of variance in cases such as ours, namely, those representing (1) differences between 'Persons',

[1] A somewhat similar assumption is adopted by Kuder and Richardson ('The Theory of the Estimation of Test Reliability', *Psychometrika*, II, 1937, pp. 151-60).

[2] M. K. Barakat, 'Factorial Study of Mathematical Ability', this *Journal*, IV, 1951, pp. 144 f. (who gives detailed references for the requisite computational procedures), and E. M. Moursy, *ibid.*, V, 1952, pp. 166 f.

[3] The relative superiority of the student or pupil in one or other of these forms of test often seemed to correspond with his intellectual type in other kinds of work. The distinction between analytic and synthetic types is not entirely new. We subsequently learnt that, when using various forms of non-verbal test (picture-completion, analogies, matrices, etc.), Burt and Moore had also noted a somewhat similar contrast: (cf. 'Experimental Tests of Higher Mental Processes', *J. Exp. Ped.*, V, p. 125, and *Brit. J. Educ. Psychol.*, XIX, p. 192).

(2) 'Interactions between Persons and Tests', and (3) 'Interaction between Persons and Trials' (or 'Occasions', as Burt prefers to say: cf. Table I). But the results of the further analysis of variance described above, and still more of the group factor analysis which we subsequently carried out, made it clear that this identification was only partly correct and might easily prove misleading.[1]

TABLE V. ANALYSIS BY SIMPLE SUMMATION: GENERAL AND BIPOLAR FACTORS

Tests	I	II	III
First Trial			
1. Picture	·924	·251	·223
2. Block	·860	·371	·142
3. Maze	·721	−·430	·317
4. Matrix	·781	−·309	·213
Second Trial			
1. Picture	·877	·291	−·154
2. Block	·804	·345	−·211
3. Maze	·803	−·340	−·185
4. Matrix	·226	−·178	−·344
Factor Variance	5·015	·831	·437
Contribution to Total Variance (per cent.)	62·7	10·4	5·5

TABLE VI. GROUP FACTOR ANALYSIS: BASIC, GROUP, AND SPECIFIC FACTORS

Tests	Basic	Anal.	Synth.	1st Day	2nd Day	Specifics		
First Trial								
1. Picture	·791	·452	·098	·341	−·045	·153		
2. Block	·689	·523	−·058	·337	·108		·312	
3. Maze	·597	−·076	·501	·412	−·064			·187
4. Matrix	·608	·101	·464	·376	·013			
Second Trial								·255
1. Picture	·827	·392	−·033	·052	·290	·153		
2. Block	·753	·431	−·008	−·106	·276		·312	
3. Maze	·701	−·112	·422	·130	·381			·187
4. Matrix	·724	·086	·387	−·076	·409			
								·255
Factor Variance	4·094	·857	·808	·578	·493	·047	·194	·070 ·130
Contribution to Total Variance (per cent.)	51·2	10·7	10·1	7·3	6·1	0·6	2·4	0·9 1·6

(b) *Group Factor Analysis.* Following Barakat we decided to carry out rotations both with Thurstone's graphical method and with Burt's arithmetical method. We began with a blind graphical procedure, rotating the summation factors two at a time in accordance with Thurstone's instructions.[2]

[1] In a personal note Professor Burt tells me that he himself would *not* make such an identification. "In the comparison printed in this Journal", he writes, "I took an extremely simple example: two tests only were applied on each day; hence, in this case both interactions *could* be represented by a single row of 'factor measurements' (cf. Table III, *loc. cit.*), and these 'factor measurements' might be expected to evince a close resemblance to those calculated from the bipolar (summational) factor analysis. In general, however, when n subtests have been applied, the residual matrix containing the interaction between the persons and the subtests will be a $2n \times N$ matrix, and will involve $(n-1)$ degrees of freedom, not one. Hence, with four subtests, the matrix would have an order of $8 \times N$, and *three* degrees of freedom, not one. Thus, unless the sample was so small as to nullify any additional factors, a single bipolar factor could not possibly account for it." And it was on his recommendation that we proceeded to undertake the group factor analysis. See also footnote (2), p. 120 above.

[2] Cf. Barakat, *loc. cit.*, p. 145. Thurstone, *Multiple Factor Analysis*, pp. 319 f.

Without changing to ' oblique ' or correlated factors, it seemed impossible to reach anything like a perfect ' simple structure ' : there were, however, several alternative structures which would reduce the variance of the general factor to about 20 or 25 per cent. of the total variance. With the ' quadrimax method ' it was reduced to 27 per cent. The reliability coefficients reached by such methods differed widely, and there seemed no definite reason for preferring one estimate to the other. At the same time there appeared to be strong grounds for rejecting all of them. Their multiplicity itself seemed to show that they were at once arbitrary and ambiguous. And above all, since the very conception of reliability implies the presence of a factor common to the various trials and the various tests, it seemed evident that procedures designed to abolish or diminish the general factor were quite unsuited to our problem.

We then attempted an arithmetical rotation, in accordance with the formulae and working methods described in this *Journal* (III, pp. 53 f.). We started by extracting one basic and four non-overlapping group factors. The lines of division between the groups were, as usual, determined by the arrangement of positive and negative saturations in the previous bipolar factors. After extracting the five factors corresponding to the significant factors in the preliminary summational analysis, we found fairly large residuals still remaining in the principal diagonal of the N.E. quadrant. These plainly indicate the presence of what would ordinarily be called ' specific factors ' peculiar to each of the subtests. The repetition of the trials converts these specific factors into narrow group-factors or ' doublets '.

Accordingly, to secure a better approximation, a fresh analysis was carried out after these quasi-specifics had been eliminated. For this purpose the self-correlations in the N.E. quadrant were first ' reduced ' by deducting the effects of the specifics, and five group factors extracted as before. To obtain saturations for the overlapping group factors a rotation-matrix was then calculated by the usual formula (*loc. cit.*, pp. 61 f.), and applied to the summational factor matrix. The results finally reached are set out in Table VI.

As usual, it was found that the ' basic factor ' accounts for rather less of the variance than the corresponding ' general factor ' obtained by the summational procedure. On the other hand, it correlates more highly with the teachers' assessments for general intelligence (0·76 instead of 0·67), and manifestly yields a better estimate of the ability our battery was designed to measure. The next two factors apparently represent specialized abilities for what we have tentatively called ' mental analysis ' and ' mental synthesis ' respectively—two irrelevant abilities which our battery was not intended to measure. The last pair of group factors represent the special conditions peculiar to one or the other of the two occasions on which the children were tested (including no doubt minor unintentional differences between the two forms of the battery).

Comparison of the Analyses. As we have seen, if we may accept the teachers' assessments of the pupils' general intelligence as a fair overall criterion, the weighted factor measurements for the ' general factor ' provided by the summational analysis would possess a higher validity than the unweighted means of the test-scores (to which the analysis of variance refers), and those for the ' basic factor ' provided by the group factor analysis a validity that is higher still.

With both methods of factorization there is clear evidence for what we could not have discovered from the analysis of variance, namely *the presence of specialized abilities for two contrasted types of test*, ' analytic ' and ' synthetic '. These specialized abilities and the effects of the two different trials seem to be represented much more clearly and more naturally by pairs of group factors than by single bipolar factors. The group factor analysis moreover discloses what the summational or bipolar analysis could scarcely be expected to reveal, namely, *the presence of quasi-specifics each peculiar to the two different forms* of the same subtest. Now that these various supplementary factors have been detected by factor analysis, we could, if we wished, go on to test their significance by a more elaborate analysis of variance (here they are fully significant) ; and accordingly it seems plain that in future investigations the factor analysis might well be undertaken at the very outset.

In the present inquiry, with the ' factorial analysis of variance ' the component for persons contributes 58 per cent. to the total variance. With the summational analysis the general factor contributes nearly 63 per cent. With the group factor analysis the basic factor contributes only 51 per cent. But, since the basic factor corresponds more closely with the ability we set out to measure, the last figure, which is also the lowest, here provides the safest guide.[1] Apparently with the analysis of

[1] Of course, had our aim been to test the combined resultant of *all* these abilities, this conclusion would no longer hold good. The analysis thus makes it clear that, as Burt puts it, " our formula for reliability must depend in part upon our specification of what the test is intended to measure. In that sense, the reliability we aim at is the reliability of the test-measurements, not merely of the test material. . . . Thus, if our aim is to measure ' innate general ability ', which by definition does not change or fluctuate, we may, with the same test-problems, get at least two alternative sets of measurements, namely, measurements expressed as multiples of the standard deviations of the age-group and measurements expressed as I.Q.'s. The latter nearly always have a lower reliability. Weighting the tests will introduce further differences. For this reason, as already emphasized, the word ' test ' is to be regarded as a shorthand term covering the whole method of measurement." These comments deserve re-quoting, because the points they bring out have apparently been overlooked by several recent critics (e.g., by A. Heim in her chapter on ' Test reliability ' in *The Appraisal of Intelligence* (1954), pp. 67 f.).

variance and with the summational analysis the results obtained were swollen by the inclusion of some of the variance attributable to the more *specialized abilities.*

With the analysis of variance the proportion attributed to the interaction *between persons and* tests was 19 per cent. This value ($\sigma^2_{pt} = \bar{r}_{sd} - \bar{r}_d$), it will be remembered, was presumed to represent the influence of specific factors tending to raise the 'self-correlations' for two applications above the 'reduced value' that we should expect if the irrelevant effects of these 'specifics' were eliminated. According to the group factor analysis, however, this estimate was excessive: for the four specifics lumped together it should apparently be nearer 5·5 per cent. The exaggerated estimate was doubtless due to the fact that, with the analysis of variance, the influence of the two group factors for mental synthesis and mental analysis (which was not separated out in the analysis of variance) has, in part at any rate, got incorporated in the interaction between persons and tests. Here again, therefore, the group factor analysis and the use of appropriate weighting seem more trustworthy.

The same holds true of the estimates for the influence of the different trials. For our present purposes this component is the most important of all, since it indicates the relative instability of the measurements obtained from virtually the same battery on different occasions. According to the analysis of variance it accounted for 11 per cent. of the total variance; according to the bipolar analysis, for only 5·5 per cent.; and according to the group factor analysis it contributes 7 per cent. on the first occasion and 6 per cent. on the second, i.e., 13 per cent. in all. In the present experiment the peculiar conditions distinguishing the two different trials are most naturally represented by two different group factors: to convert these two group factors into a single bipolar factor (as is done explicitly by the summational analysis and implicitly by the analysis of variance) may easily distort the facts and underestimate the influence of such disturbing conditions. All these arguments, therefore, point to the same conclusion, namely, that, for such problems as the present, *the factorial analyses are superior to the analysis of variance*, and of the two factorial procedures *the group factor method is superior to the bipolar or summational.*

Final Assessment of Reliability. From the foregoing results one practical corollary stands out in sharp relief: before the investigator can estimate the reliability of his test, or even decide on an adequate procedure for so doing, he must first settle the question—reliability for what? Here the answer is quite plain: our tests were constructed to measure 'general intelligence'; and, from the evidence available, there can be little doubt that our best approximation to this is to be found in the 'basic factor' common to all the tests.

To obtain the best assessments for this basic factor, the obvious procedure is to compute the regression coefficients by the usual formula $f'R^{-1}$, where f' is the row-vector of saturations for the basic factor (Table VIII, col. 1) and R^{-1} the inverse of the relevant correlation matrix (Table II). Then, instead of taking the unweighted sum of each pupil's marks for each subtest (p. 124) we shall use the regression coefficients to weight the marks obtained at the first and the second trial respectively.[1]

To obtain an improved estimate for reliability based on the factor measurements, the requisite equation is similar in form to that already given (eqn. ii); but the row vector w' will now consist of differential weights (the regression coefficients) instead of equal weights. The reliability coefficient thus computed proves to be 0·838. This figure therefore provides the most appropriate estimate of reliability for our present purpose.[2]

We should, however, like to insist that the determination of a single final figure for reliability, however ingeniously reached, is by no means the sole or the main conclusion to be derived from such analyses as the present. As the reader will have perceived, the alternative and incidental calculations described in the foregoing sections have given that figure a clearer meaning, and at the same time have greatly increased our detailed understanding of the tests and of the points at which they need to be revised. Here, however, our object was not to report any specific results achieved by this particular inquiry, but merely to use our data to illustrate what we consider to be improvements in current statistical techniques.

[1] These regressions are themselves nearly always instructive. Here, for example, small negative weights are found for the Block and Matrix tests, which thus operate as 'suppressor-variables', and counteract the effects of the two irrelevant specialized abilities. The Picture test has the highest weight, and would seem to indicate the most reliable type of subtest for a population of this kind. We are therefore hoping to improve the efficiency of the other tests by reconstructing them with more colourful and pictorial material.

[2] It might have been expected that the introduction of negative weights would reduce the value now obtained for the reliability coefficient below that previously reached with equalized weights. With equalized weights, the figure for reliability would (one might suppose) be increased by the inclusion of the more specialized abilities. That influence is certainly traceable; but its removal seems to be more than counterbalanced by other improvements in the estimate which the differential weighting has produced. The subtests which have an intrinsically poor reliability (the Matrix test for example) now receive a very low weight. In other investigations my co-workers and I have frequently found that the use of negative weights and the elimination of irrelevant abilities *do* reduce the reliability coefficient, though they nearly always increase the validity coefficient (i.e., the multiple correlation with the general or basic factor). At times therefore—indeed, it would seem, more often than not—the use of the simpler formula for the test-retest coefficient may decidedly exaggerate the true reliability; and for this reason we believe that, in all cases of doubt, it is essential to use a group factor analysis to estimate reliability as well as validity.

INDEX TO VOLUME VIII

INDEX OF AUTHORS

Authors	*Titles*	*Page*
Burt, C.	Sir Godfrey Thomson	1–2
Burt, C.	Test Reliability Estimated by Analysis of Variance	103–118
Burt, C., Cooper, W. F., and Martin, J. L.	A Psychological Study of Typography	29–57
Cattell, R. B., and Cattell, A. K. S.	Factor Rotation for Proportional Profiles	83–92
Cureton, E. E.	On the Use of Burt's Formula for Estimating Factor Significance	28
Guttman, L.	An Additive Metric from all the Principal Components of a Perfect Scale	17–29
Guttman, L.	The Determinacy of Factor Score Matrices with Implications for Five Other Basic Problems of Common-Factor Theory	65–81
Mahmoud, A. F.	Test Reliability in Terms of Factor Theory	119–135
Roberts, J. A. F., and Dunsdon, M. I.	A Study of the Performance of 2,000 Children on Four Vocabulary Tests	3–16
Simon, H. A., and Guetzkow, H.	Mechanisms Involved in Group Pressures on Deviate-Members	93–101
Stuart, A.	The Correlation between Variate Values and Ranks	25–27

INDEX OF SUBJECTS

Titles	*Authors*	*Page*
Factor Score Matrices, the Determinacy of, with Implications for Five Other Basic Problems of Common-Factor Theory	Guttman, L.	65–81
Factor Significance, On the Use of Burt's Formula for Estimating	Cureton, E. E.	28
Group Pressures on Deviate Members, Mechanisms involved in	Simon, H. A. and Guetzkow, H	93–101
Principal Components of a Perfect Scale, An Additive Metric from all the	Guttman, L.	17–29
Proportional Profiles, Factor Rotation for	Cattell, R. B., and Cattell, A. K. S.	83–92
Ranks, The Correlation between Variate Values and	Stuart, A.	25–27
Test Reliability Estimated by Analysis of Variance	Burt, C.	103–118
Test Reliability in Terms of Factor Theory	Mahmoud, A. F.	119–135
Thomson, Sir Godfrey	Burt, C.	1–2
Typography, A Psychological Study of	Burt, C., Cooper, W. F., and Martin, J. L.	29–57
Vocabulary Tests, A Study of the Performance of 2,000 Children on Four	Roberts, J. A. F. and Dunsdon, M. I.	3–16

BOOKS REVIEWED

Author	*Title*	*Page*
Bowden, B. V. (ed.)	Faster than Thought: A Symposium on Digital Computing Machines	59–64


EDUC. & PSYL. RESEARCH
BUREAU OF
JUL 1962